Статистика – 3. Дескриптивный корреляционный и регрессионный анализ

Д-р Вольфганг Бёме

Декабрь 2010

Содержание

- 1. Введение
- 2. Дескриптивный одномерный анализ распределений
- **3. Дескриптивный корреляционный и регрессионный анализ**
- 4. Дескриптивный анализ временных рядов
- 5. Относительные показатели / Индексы
- 6. Введение в избранные многомерные методы анализа

- Статистическое программное обеспечение: SPSS

Постановка задачи

Несколько метрически масштабированных признаков X^1, X^2,...,X^K

Первичные списки для каждого метрически масштабированного признака:

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

Можно ли статистически доказать наличие связи между признаками? $\Rightarrow$ **Корреляционный анализ**

Вывод математической формулы для описания связи, напр.:

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

$\Rightarrow$ **Регрессионный анализ**

Ограничение: пока $K = 2 \Rightarrow$ только два признака

2. Корреляционный и регрессионный анализ – Содержание

- **3.1 Корреляционный анализ**
- 3.2 Простая линейная регрессия
- 3.3 Простая нелинейная регрессия
- 3.4 Множественная регрессия

Содержание

- **3.1.1 Двумерное распределение частот**
- 3.1.2 Контингентный анализ
- 3.1.3 Условные частоты / Независимость
- 3.1.4 Метрический корреляционный анализ
- 3.1.5 Ранговый корреляционный анализ
- 3.1.6 Множественный корреляционный анализ

Двумерное распределение частот

Пусть X и Y – два метрически масштабированных признака с двумя первичными списками:

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

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

или двумерный первичный список:

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

Двумерное распределение частот

Рассматриваем несовпадающие значения признаков (ξ_l, $l = 1, 2, ..., m_1$ и ν_r, $r = 1, 2, ..., m_2$) обоих признаков и соответствующие абсолютные / относительные частоты n_{lr}/p_{lr}. Тогда можно образовать следующие пары данных:

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

Где обозначают:

$$\begin{aligned} n_{lr} \quad &- \quad \text{число появлений } (\xi_l, \nu_r) \\ p_{lr} \quad &= \quad \frac{n_{lr}}{n} \end{aligned}$$

Пример – данные

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – 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_j	r	v_r
1	6.0	1	125
2	6.3	2	135
3	6.4	3	140
4	6.6		
5	6.7		

Пример – одномерные распределения частот

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – 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

Двумерное распределение частот

Совместное двумерное распределение частот

Определение: Абсолютное распределение частот – это соответствие несовпадающих значений признака их абсолютным частотам:

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

Определение: Относительное распределение частот – это соответствие несовпадающих значений признака их относительным частотам:

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

Пример – двумерное абсолютное распределение частот

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – 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

Пример – двумерное относительное распределение частот

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y

mehrdimensionale Häufigkeitstabellen

		r	1	2	3
	X / Y		v_1	v_2	v_3
I			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

Двумерное распределение частот

Маргинальные частоты

Определение:

Абсолютные маргинальные частоты признака X и Y:

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

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

Двумерное распределение частот

Определение:

Относительные маргинальные частоты признака X и Y:

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

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

Двумерное распределение частот

Маргинальные распределения

Определение: Абсолютное маргинальное распределение – это соответствие несовпадающих значений признаков X и Y их абсолютным маргинальным частотам:

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

Определение: Относительное маргинальное распределение – это соответствие несовпадающих значений признаков X и Y их относительным маргинальным частотам:

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

Пример – абсолютные маргинальные частоты

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y

Randverteilung 1

			n_{i*}
1	ξ1	6	2
2	ξ2	6.3	2
3	ξ3	6.4	1
4	ξ4	6.6	2
5	ξ5	6.7	3

Randverteilung 2

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

Содержание

- 3.1.1 Двумерное распределение частот
- **3.1.2 Контингентный анализ**
- 3.1.3 Условные частоты / Независимость
- 3.1.4 Метрический корреляционный анализ
- 3.1.5 Ранговый корреляционный анализ
- 3.1.6 Множественный корреляционный анализ

Таблица сопряжённости

	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

Пример – абсолютная таблица сопряжённости

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y

Kontingenztabelle

		r	1	2	3	
	X / Y		v_1	v_2	v_3	n_{l*}
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

Таблица сопряжённости с группировкой по интервалам

	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

Пример

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y
(с группировкой по интервалам)

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

I	I_l	r	I_r
1	bis 6,00	1	bis 135
2	6,01bis 6,8	2	>135
3	> 6,60		

Пример

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y
(с группировкой по интервалам)

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

Пример

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y
(с группировкой по интервалам)

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

Содержание

- 3.1.1 Двумерное распределение частот
- 3.1.2 Контингентный анализ
- **3.1.3 Условные частоты / Независимость**
- 3.1.4 Метрический корреляционный анализ
- 3.1.5 Ранговый корреляционный анализ
- 3.1.6 Множественный корреляционный анализ

Условные частоты

Условные частоты: рассмотрение одного признака при фиксировании другого на некотором значении

Определение: Условные частоты признака X при фиксированном Y:

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

Определение: Условные частоты признака Y при фиксированном X:

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

Определение: **Относительное условное распределение** – соответствие несовпадающих значений признаков X и Y их условным частотам:

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

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

Пример – относительная таблица сопряжённости

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – 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	

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

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

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

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

Пример – условные частоты

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – 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

Независимость

Исходная точка – условные частоты: рассмотрение одного признака при фиксировании другого на некотором значении

Определение: Признаки X и Y называются **независимыми** тогда и только тогда, когда выполняется:

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

Пример

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y

Независимость

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

Содержание

- 3.1.1 Двумерное распределение частот
- 3.1.2 Контингентный анализ
- 3.1.3 Условные частоты / Независимость
- **3.1.4 Метрический корреляционный анализ**
- 3.1.5 Ранговый корреляционный анализ
- 3.1.6 Множественный корреляционный анализ

Эмпирическая ковариация

Определение: Пусть X и Y – два метрически масштабированных признака. Значения этих двух признаков получены у n статистических единиц:

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

Тогда эмпирическая ковариация d_{XY} определяется следующим образом:

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

Эмпирическая ковариация

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

поскольку

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

Свойства эмпирической ковариации

- d_{XY} измеряет силу линейной связи между признаками X и Y.
- Относительно знака:
  - $d_{XY} < 0$ – обратная связь
  - $d_{XY} = 0$ – нет связи
  - $d_{XY} > 0$ – прямая связь
- d_{XY} имеет размерность

Свойства эмпирической ковариации – продолжение

- Симметричная характеристика (матрица дисперсий-ковариаций)

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

где

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

- Линейное преобразование: $\widetilde{x}_i = \alpha_1 + \beta_1 x_i$ и $\widetilde{y}_i = \alpha_2 + \beta_2 y_i$

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

Коэффициент корреляции

Определение: Пусть X и Y – два метрически масштабированных признака. Значения получены у n статистических единиц:

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

Тогда **эмпирический простой линейный коэффициент корреляции** r_{XY} определяется:

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

Синонимы:

- Бивариантный или двойной коэффициент корреляции
- Метрический коэффициент корреляции по БРАВЭ и ПИРСОНУ

Свойства коэффициента корреляции

- r_{XY} измеряет силу линейной связи между признаками X и Y. Область значений: $-1 \leq r_{XY} \leq 1$
- Выполняется:

$$
\begin{aligned}
r_{XY} = 1 &\quad \Rightarrow \quad \text{полная положительная корреляция} \\
0 \leq r_{XY} \leq 1 &\quad \Rightarrow \quad \text{положительная корреляция} \\
r_{XY} = 0 &\quad \Rightarrow \quad \text{некоррелированность} \\
-1 \leq r_{XY} \leq 0 &\quad \Rightarrow \quad \text{отрицательная корреляция} \\
r_{XY} = -1 &\quad \Rightarrow \quad \text{полная отрицательная корреляция}
\end{aligned}
$$

- r_{XY} безразмерен
- Нормированная характеристика

Свойства коэффициента корреляции – продолжение

- Симметричная характеристика (матрица корреляций)

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

где

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

- Линейное преобразование:

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

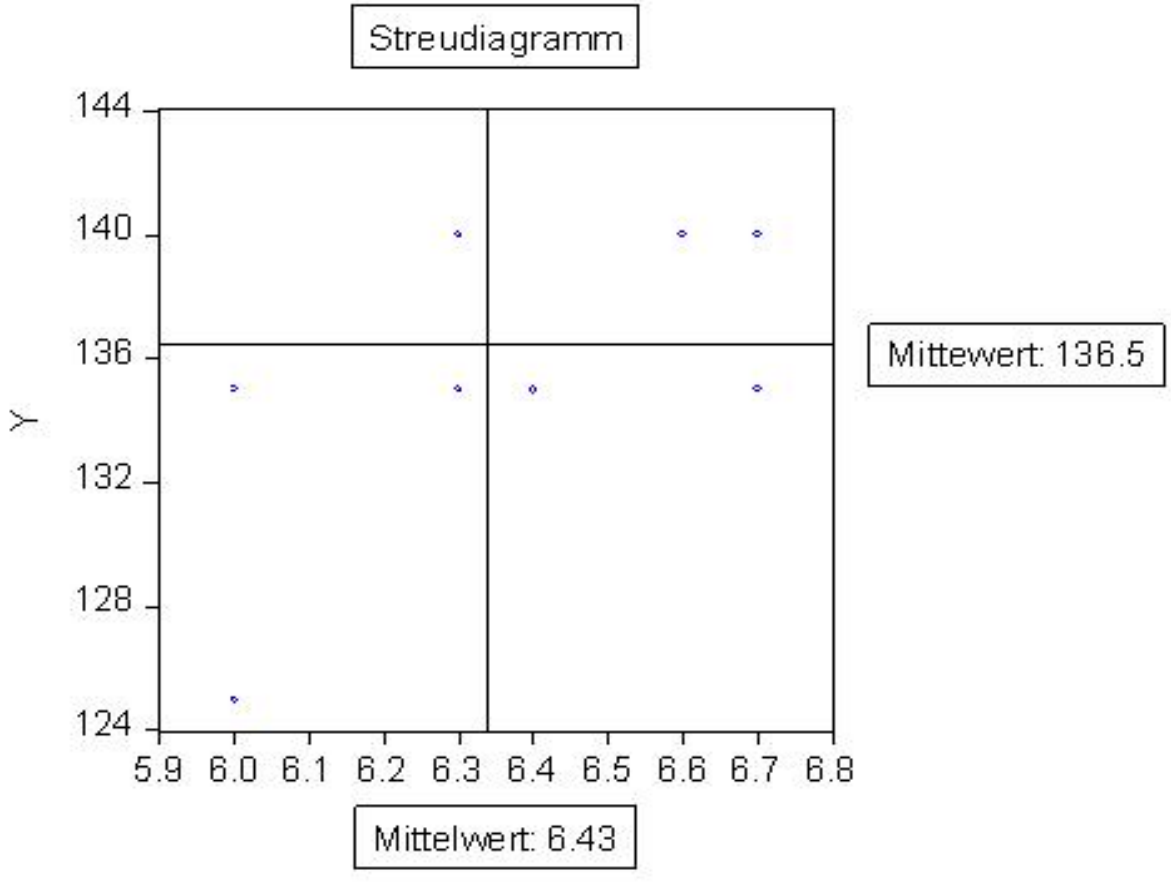
Пример – корреляция / эмпирическая ковариация

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

$$
\begin{aligned}
d_{XY} &= \overline{xy} - \overline{x} \cdot \overline{y} = 878{,}5 - 6{,}4 \cdot 136{,}5 = 0{,}805 \\
d_x^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = 0{,}07 \\
d_y^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2 = 20{,}3 \\
r_{XY} &= \frac{d_{XY}}{\sqrt{d_X^2}\sqrt{d_Y^2}} = \frac{0{,}805}{\sqrt{0{,}07}\sqrt{20{,}3}} = \frac{0{,}805}{1{,}174} = 0{,}686
\end{aligned}
$$

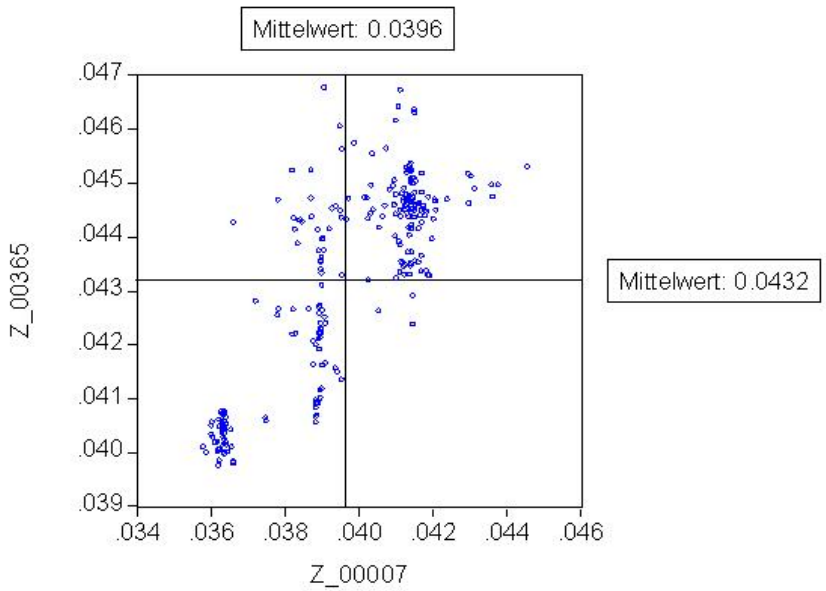
Корреляция / эмпирическая ковариация – графическая интерпретация

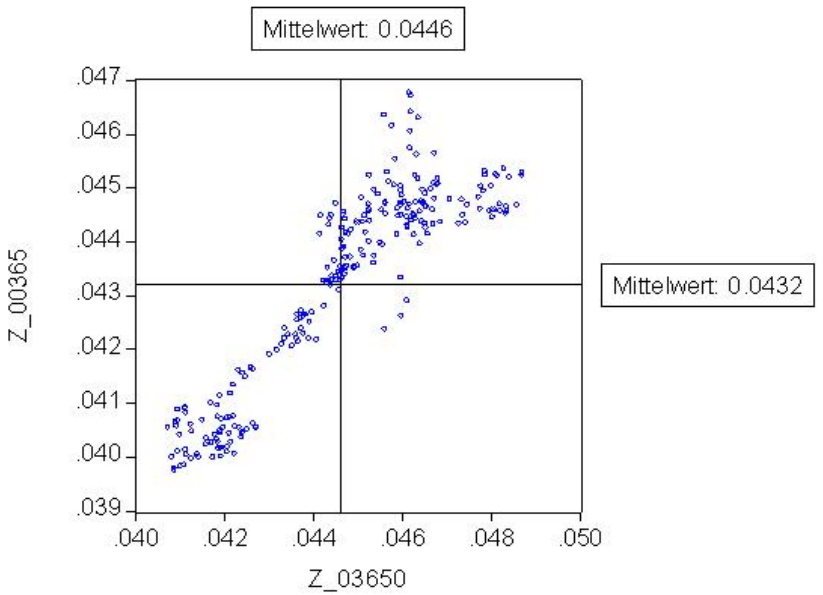
Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y



Корреляция / эмпирическая ковариация – графическая интерпретация

Пример: процентные ставки 2007





Матрица корреляций / матрица дисперсий-ковариаций

Пример: процентные ставки 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

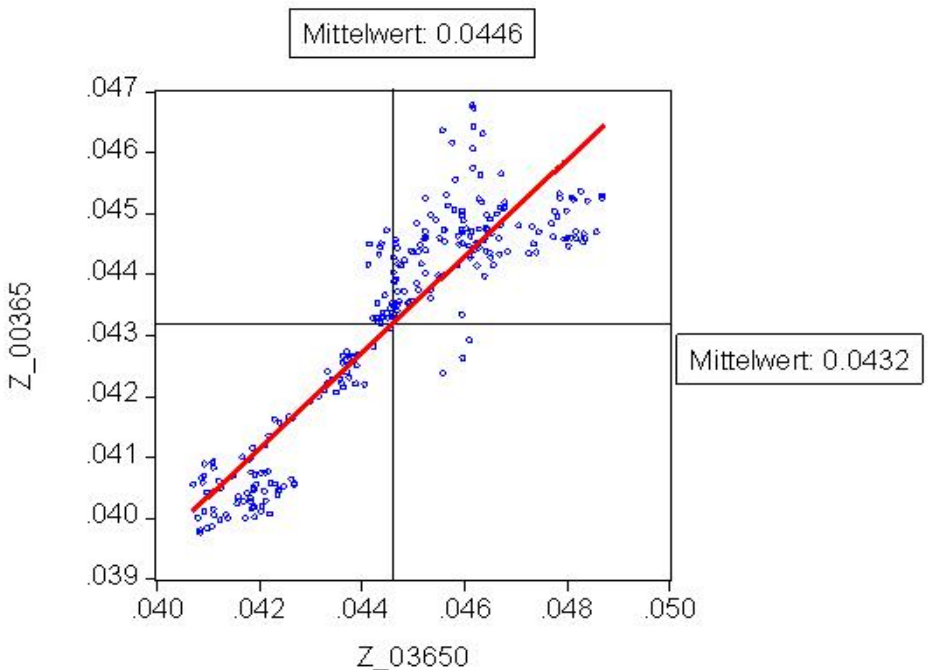
	Z_00007	Z_00365
Z_00007	0.0000043	0.0000032
Z_00365	0.0000032	0.0000035

Varianz-Kovarianzmatrix

	Z_03650	Z_00365
Z_03650	0.0000046	0.0000036
Z_00365	0.0000036	0.0000035

Корреляция / линейная связь – графическая интерпретация

Пример: процентные ставки 2007



$\Longrightarrow$ см. 3.2 Простая линейная регрессия

Введение в SPSS – темы

Статистические темы: дескриптивная статистика

- Частоты – продолжение
- Дескриптивная статистика – продолжение
- Бивариантные корреляции

Метрическая корреляция – пример

с Excel:

Пример 1: BeispielW95.xls
Пример 2: Zinssaetze 2007_2.xls

Задание: определить ковариации, коэффициенты корреляции и матрицу ковариаций

с SPSS:

Пример 1: World95.sav.xls
Пример 2: Zinssaetze 2007_2.xls

Задание: определить коэффициенты корреляции

Содержание

- 3.1.1 Двумерное распределение частот
- 3.1.2 Контингентный анализ
- 3.1.3 Условные частоты / Независимость
- 3.1.4 Метрический корреляционный анализ
- **3.1.5 Ранговый корреляционный анализ**
- 3.1.6 Множественный корреляционный анализ

Ранговая корреляция

Определение: Пусть X – хотя бы порядково масштабированный признак. Для признака имеется ранговый ряд порядкового набора данных:

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

Тогда порядковый номер i, присвоенный значению признака x_i, называется ранговым числом и обозначается R_i^X.

Проблема: одинаковые значения признака (**связанные ранги**).

SPSS: среднее арифметическое, макс., мин., сквозная нумерация

Здесь:

среднее арифметическое $\Rightarrow \frac{n\cdot(n+1)}{2} = \sum_{i=1}^{n} R_i^X$ или

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

Ранговая корреляция

Определение: Пусть X и Y – два хотя бы порядково масштабированных признака. Значения получены у n статистических единиц:

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

Каждой паре значений (x_i, y_i) соответствует пара ранговых чисел (R_i^X, R_i^Y). Тогда величина:

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

называется **ранговым коэффициентом корреляции СПИРМЕНА**.

Ранговая корреляция – интерпретация

- r_S измеряет силу линейной связи между признаками X и Y. Область значений: $-1 \leq r_S \leq 1$
- Относительно знака:
  - $r_S < 0$ – обратная связь
  - $r_S = 0$ – нет связи
  - $r_S > 0$ – прямая связь
- Если r_S близко к нулю – связи нет; если r_S близко к 1 или -1 – есть прямая или обратная связь.

Ранговая корреляция – пример

Пример: расход бензина / выброс загрязняющих веществ 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

$$r_S = 1 - \frac{6 \cdot 42}{10(100-1)} = 0{,}745$$

Темы SPSS

Статистические темы: дескриптивная статистика

- Бивариантные корреляции – продолжение

Ранговая корреляция – пример

с Excel:

Пример 1: BeispielW95.xls
Пример 2: Zinssaetze 2007_2.xls

Задание: определить ранговые коэффициенты корреляции

с SPSS:

Пример 1: World95.sav.xls
Пример 2: Zinssaetze 2007_2.xls

Задание: определить ранговые коэффициенты корреляции

Содержание

- 3.1.1 Двумерное распределение частот
- 3.1.2 Контингентный анализ
- 3.1.3 Условные частоты / Независимость
- 3.1.4 Метрический корреляционный анализ
- 3.1.5 Ранговый корреляционный анализ
- **3.1.6 Множественный корреляционный анализ**

Постановка задачи

Ранее: X и Y – два метрически масштабированных признака с двумя первичными списками:

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

или двумерный первичный список: $(x_1, y_1),\ (x_2, y_2),\ \ldots, (x_n, y_n)$

Теперь: X_j, $j = 1, 2, \ldots, m$ – метрически масштабированные признаки с m первичными списками:

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

или m-мерный первичный список:

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

Эмпирические ковариации

Определение: X_j, $j = 1, 2, ..., m$ – метрически масштабированные признаки. Значения получены у n статистических единиц. Тогда эмпирическая ковариация d_{jk} определяется:

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

Матрица ковариаций:

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

Эмпирические коэффициенты корреляции

Определение: X_j, $j = 1, 2, ..., m$ – метрически масштабированные признаки. Тогда коэффициент корреляции r_{jk}:

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

Матрица корреляций:

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

Введение в SPSS – темы

Статистические темы: дескриптивная статистика

- Бивариантные корреляции – продолжение

Множественная корреляция – пример

с Excel:

Пример 2: Zinssaetze 2007_2.xls

Задание: определить матрицу корреляций и матрицу ковариаций

с SPSS:

Пример 2: Zinssaetze 2007_2.xls

Задание: определить коэффициенты корреляции

2. Корреляционный и регрессионный анализ – Содержание

- 3.1 Корреляционный анализ
- **3.2 Простая линейная регрессия**
- 3.3 Простая нелинейная регрессия
- 3.4 Множественная регрессия

Содержание

- **3.2.1 Постановка задачи и вывод**
- 3.2.2 Интерпретация и пример
- 3.2.3 Эластичности
- 3.2.4 Меры качества

Постановка задачи

Определение: Пусть X и Y – два метрически масштабированных признака. Значения получены у n статистических единиц:

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

Линейная функция $y^* = f(x) = b_0 + b_1 \cdot x$ для описания статистической зависимости признака Y от признака X называется **простой линейной регрессией** Y на X.

Постановка задачи

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

Обозначения:

X – регрессор, независимая или экзогенная переменная
Y – регрессант, зависимая или эндогенная переменная
x – реализация признака X ($x \in \mathbb{R}$)
y – реализация признака Y ($y \in \mathbb{R}$)
$y^* = f(x) = b_0 + b_1 \cdot x$ – регрессионное значение (реализация y при заданном x)
b_0 – константа регрессии (intercept)
b_1 – линейный коэффициент регрессии (slope)
$u = y - y^*$ – ошибка (остаток)

Постановка задачи

Линейная регрессионная функция: $y^* = f(x) = b_0 + b_1 \cdot x$, $x \in \mathbb{R}$

x – наблюдаемое значение признака X
$y^* = y^*(x)$ – вычисленное (расчётное) значение признака Y
y – наблюдаемое значение признака Y

$$\begin{aligned} u_i &= y_i - y_i^* \text{ остаток (ошибка)} \\ y_i &= y_i^* + u_i = b_0 + b_1 \cdot x_i + u_i \end{aligned}$$

Требование:

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

$\Rightarrow$ **Метод наименьших квадратов**

Вывод

Вывод **метода наименьших квадратов**

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

Задача на экстремум функции двух переменных (b_0, b_1)

Необходимое условие:

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

Вывод – продолжение

Необходимое условие:

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

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

Вывод – продолжение

Система уравнений:

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

Решение системы уравнений

$$
\begin{aligned}
b_0 &= \overline{y} - b_1 \overline{x} \\
b_1 &= \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{d_{YX}}{d_X^2}
\end{aligned}
$$

Вывод – продолжение

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

Достаточное условие:

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

Вывод матричная запись

Уравнение $y = f(x) = b_0 + b_1 \cdot x + u$ выполняется для всех наблюдений первичного списка:

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

или в матричной записи:

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

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

Операции с матрицами

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

Вывод

Исходная точка: $y = X \cdot b + u$
Правила: $(A + B)^T = A^T + B^T$ и $(AB)^T = B^T A^T$

Задача на экстремум:

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

Система нормальных уравнений:
$X^T Xb = X^T y \implies$

Решение:

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

Содержание

- 3.2.1 Постановка задачи и вывод
- **3.2.2 Интерпретация и пример**
- 3.2.3 Эластичности
- 3.2.4 Меры качества

Реалистичный пример

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):

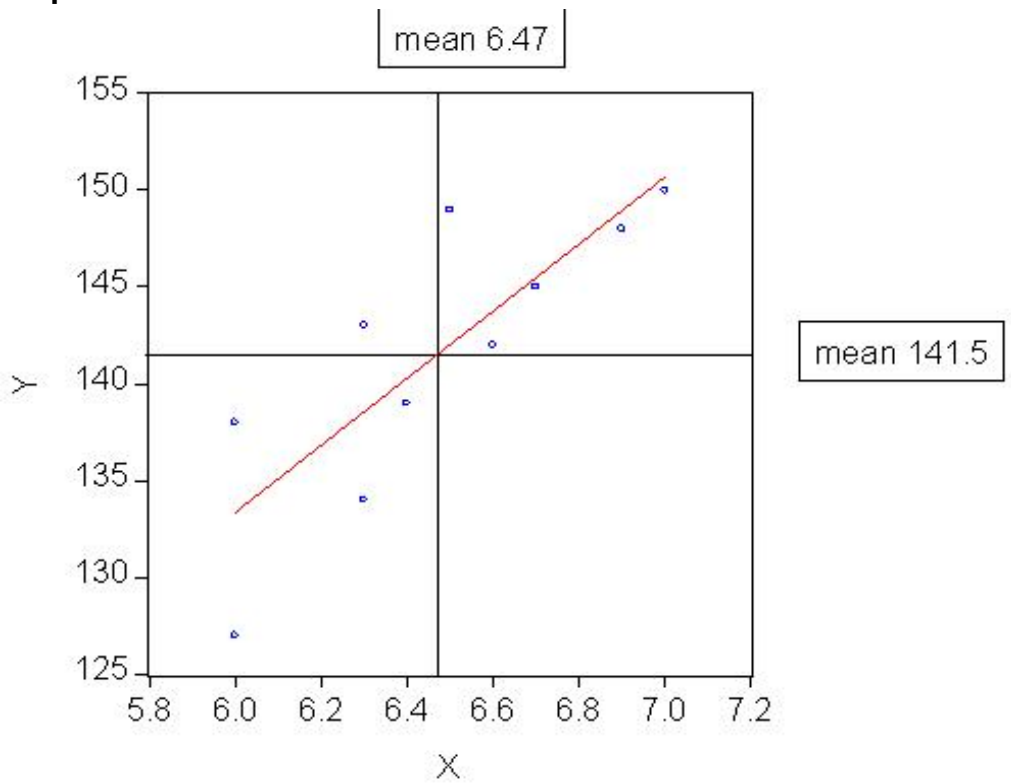
PKW-Nummer	Kraftstoffverbrauch l / 100 km	Kohlendioxidausstoß g / km	Fabrikat / Modell
Sp. 1	Sp. 2	Sp. 3	Sp. 4
1	6,0	138	Opel Corsa
2	6,0	127	Daihatsu Cuore
3	6,3	143	Renault Twingo
4	6,3	134	Suzuki Alto
5	6,4	139	Seat Arosa
6	6,5	149	Peugot 106
7	6,6	142	VW Polo
8	6,7	145	Fiat Cinquecento 0.9 i.e
9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

Графическая интерпретация

Пример: расход топлива (л/100 км) – X и выброс загрязняющих веществ (г/км) – Y



Вычисления

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

$$b_0 = \frac{1415{,}0 \cdot 419{,}7 - 64{,}7 \cdot 9173{,}1}{10 \cdot 419{,}7 - 4186{,}09} = 29{,}31$$

$$b_1 = \frac{10 \cdot 9173{,}1 - 64{,}7 \cdot 1415{,}0}{10 \cdot 419{,}7 - 4186{,}09} = 17{,}34$$

Интерпретация

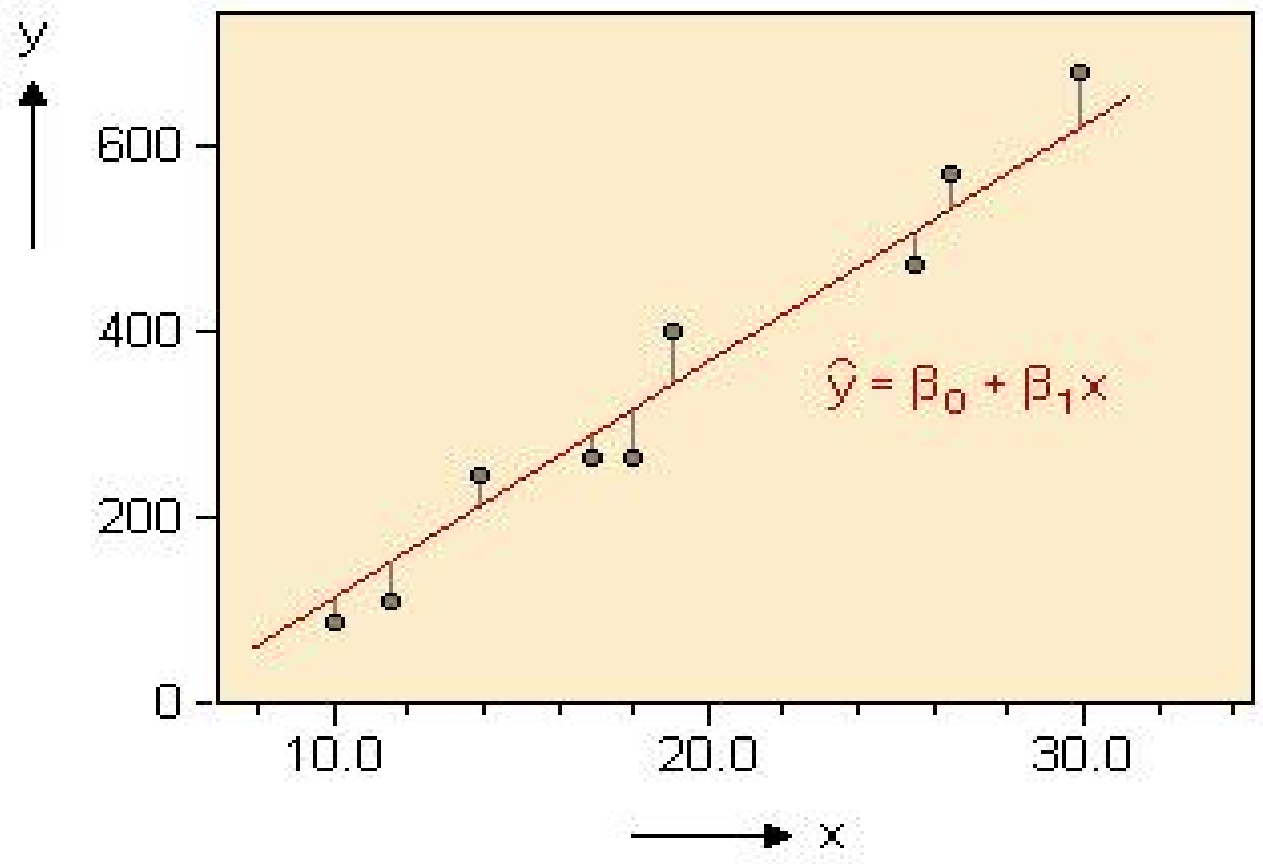
Единицы измерения: X – расход топлива [$\frac{}{100}$], Y – выброс загрязнений [ ]

$y^* = 29{,}31 + 17{,}34 \cdot x$

Интерпретация коэффициентов b_0 и b_1:

- b_0 – если $x = 0$ (нет расхода топлива!), тем не менее выброс составляет 29,31 [ ]. $\Longrightarrow$ Выброс, не зависящий от расхода топлива.
- b_1 – при расходе топлива на 1 $\frac{}{100}$ выброс составляет 17,34 , или при увеличении / уменьшении расхода на 1 $\frac{}{100}$ выброс растёт / снижается на 17,34 .

Графическая интерпретация



Свойства линейной регрессии

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

Содержание

- 3.2.1 Постановка задачи и вывод
- 3.2.2 Интерпретация и пример
- **3.2.3 Эластичности**
- 3.2.4 Меры качества

Предельная функция / абсолютное изменение

Определение: Пусть $y^*(x)$ – непрерывная бивариантная регрессионная функция. Тогда первая производная $y'^*(x)$ называется **предельной функцией регрессионной функции**:

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

Значение в точке x_0 называется **среднеабсолютным изменением** или **среднеабсолютной эластичностью**.
Если регрессионная функция линейна ($y^* = b_0 + b_1 \cdot x$), то:

$$y'*(x_0) = b_1 \ \forall x = x_0$$

Эластичность / функция эластичности

Определение: функция $\epsilon(x)$ называется **функцией эластичности регрессионной функции**:

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

Значение в точке x_0 – **эластичность переменной Y** или **среднеотносительное изменение**.
Если $y^* = b_0 + b_1 \cdot x$:

$$\epsilon(x) = \frac{b_1 \cdot x}{b_0 + b_1 \cdot x}$$

Эластичность / интерпретация

Оценка реакции переменной Y на изменение X:

$$\begin{array}{lll} |\epsilon(x_0)| & > 1 & \text{сверхпропорционально эластично} \\ |\epsilon(x_0)| & = 1 & \text{пропорционально эластично} \\ |\epsilon(x_0)| & < 1 & \text{недопропорционально эластично} \end{array}$$

$y^* = 29{,}31 + 17{,}34 \cdot x$

$$\begin{aligned} y'*(x_0) &= b_1 = 17{,}34 \ \forall x = x_0 \\ \epsilon(x_0) &= \frac{17{,}34 \cdot 6{,}5}{29{,}31 + 17{,}34 \cdot 6{,}5} = 0{,}7936 \text{ при } x_0 = 6{,}5 \\ &\Longrightarrow |\epsilon(x_0)| < 1 \Longrightarrow \text{недопропорционально} \end{aligned}$$

Эластичность / интерпретация

$y^* = 29{,}31 + 17{,}34 \cdot x$

- $y'*(x_0)$ – **среднеабсолютное изменение переменной Y**
  при увеличении/уменьшении расхода топлива на 1 $\frac{}{100}$ выброс растёт/снижается на 17,34 .
- $\epsilon(x_0)$ – **эластичность переменной Y**
  при увеличении/уменьшении расхода топлива на 1 процент выброс растёт/снижается на 0,79 процента.

Содержание

- 3.2.1 Постановка задачи и вывод
- 3.2.2 Интерпретация и пример
- 3.2.3 Эластичности
- **3.2.4 Меры качества**

Дисперсия / стандартное отклонение остатков

Определение: Дисперсия остатков (среднеквадратическая ошибка регрессии) d_u^2:

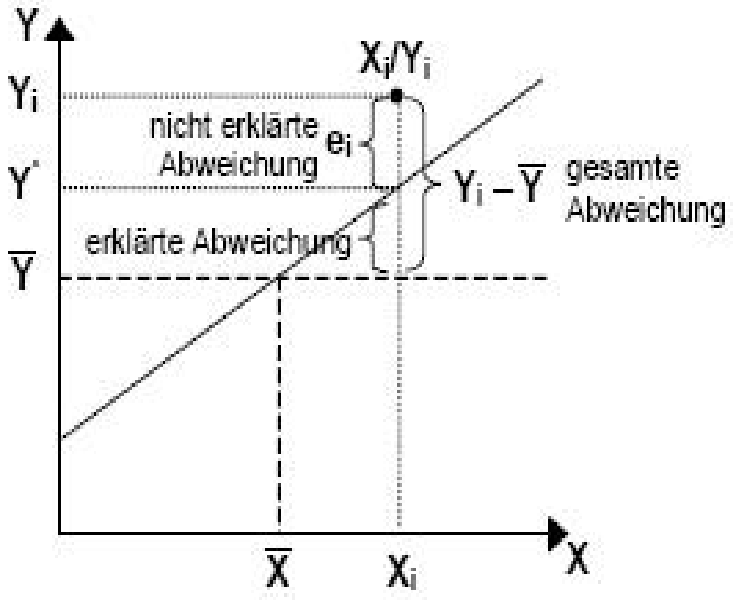
$$d_u^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$$

Определение: Стандартное отклонение остатков:

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

Коэффициент детерминации

Интерпретация:



$$R^2 = \frac{\text{объяснённое отклонение}}{\text{общее отклонение}}$$

Коэффициент детерминации

$$R^2 = \frac{\sum_{i=1}^{n}(y_i^* - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{d_{y^*}^2}{d_y^2} = 1 - \frac{d_u^2}{d_y^2}$$

Требования к мерам качества

Чтобы использовать меры качества для сравнения различных моделей, должны выполняться условия:

1. Эндогенная переменная в сравниваемых моделях должна быть идентичной (числовые значения должны совпадать).
2. Число экзогенных параметров в сравниваемых моделях должно быть одинаковым.
3. Модели имеют параметр уровня (b_0).

Введение в SPSS – темы

Статистические темы: дескриптивная статистика

- Линейная регрессия
- Диаграммы (последовательность) – продолжение

Линейная регрессия – пример

с Excel:

Пример 2: Zinssaetze 2007_2.xls

Задание: определить уравнение регрессии и меры качества

с SPSS:

Пример 2: Zinssaetze 2007_2.xls

Задание: оценить коэффициенты, анализ ANOVA и других таблиц, графическое представление результатов

2. Корреляционный и регрессионный анализ – Содержание

- 3.1 Корреляционный анализ
- 3.2 Простая линейная регрессия
- **3.3 Простая нелинейная регрессия**
- 3.4 Множественная регрессия

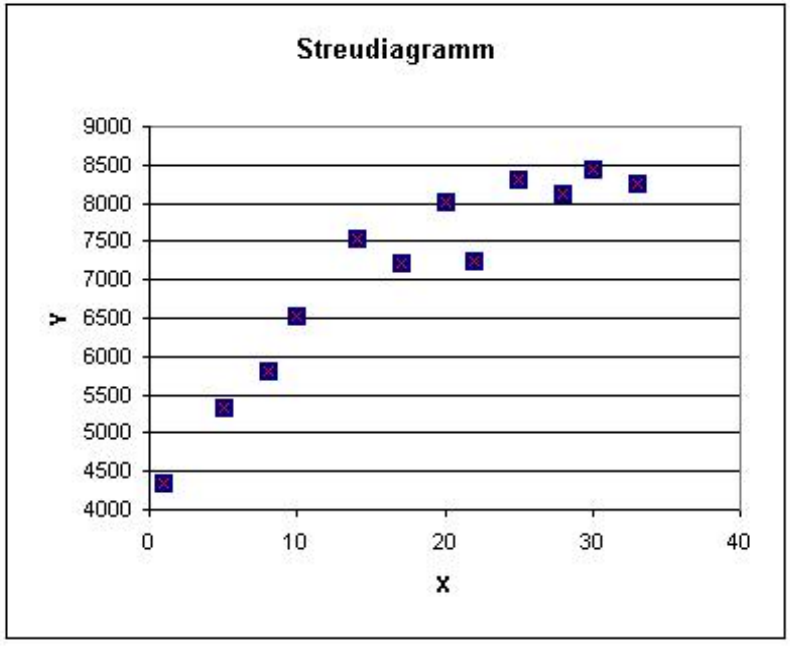
Содержание

- **3.3.1 Постановка задачи**
- 3.3.2 Линеаризация
- 3.3.3 Эластичности

Вводный пример

Пример: концентраты (в 50 кг/год) X и молочная продуктивность (л/год) Y за 12 лет

	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



Источник: von Auer, Эконометрика, Берлин, Гейдельберг, Нью-Йорк, 1999

Содержание

- 3.3.1 Постановка задачи
- **3.3.2 Линеаризация**
- 3.3.3 Эластичности

Линеаризуемые функции

Преобразование нелинейных функций в линейные: линеаризация

Тип функции	**Уравнение функции**	**Линеаризованная функция**
логарифмическая	$e^y = e^{(b_0+u)}x^{b_1}$ $y = b_0 + b_1 ln(x) + u$	$y = b_0 + b_1 ln(x) + u$
экспоненциальная	$y = e^{b_0+b_1 x+u}$	$ln(y) = b_0 + b_1 x + u$
обратная	$y = b_0 - b_1 \frac{1}{x} + u$	$y = b_0 + b_1 z + u,\ z = \frac{1}{x}$
степенная	$y = b_0 x^{b_1} e^u$	$ln(y) = ln(b_0) + b_1 ln(x) + u$

Линеаризуемые функции

Используемые преобразованные величины:

Тип функции	**Линеаризованная функция**	**Используемые наблюдения**
логарифмическая	$y = b_0 + b_1 ln(x) + u$	y, $ln(x)$
экспоненциальная	$ln(y) = b_0 + b_1 x + u$	$ln(y)$, x
обратная	$y = b_0 + b_1 z + u$, $z = \frac{1}{x}$	y, $\frac{1}{x}$
степенная	$ln(y) = ln(b_0) + b_1 ln(x) + u$	$ln(y)$, $ln(x)$

Проблемы

Техника линеаризации порождает следующие проблемы:

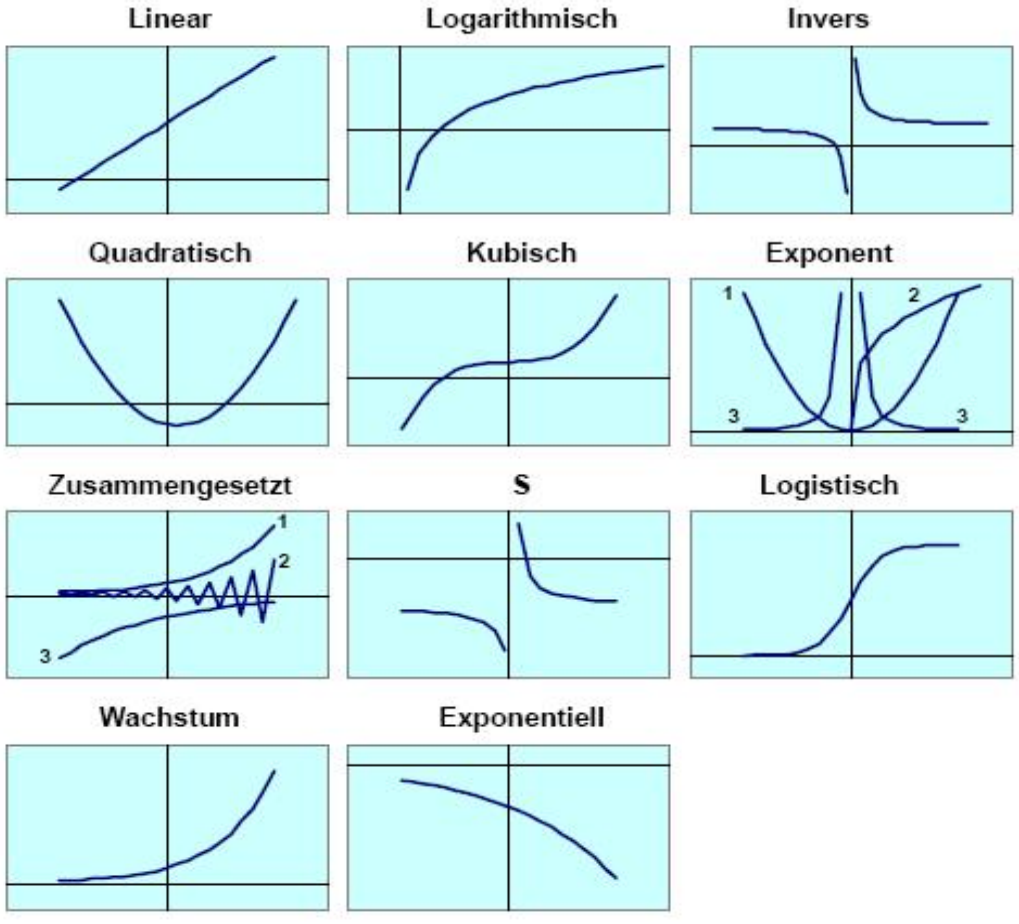
- Ошибки u имеют разные единицы измерения для каждой линеаризованной функции. Поэтому сравнение качества модели с помощью дисперсии остатков и коэффициента детерминации недопустимо.
- Не все нелинейные функции поддаются линеаризации.
- Необходимо проверить, сохраняются ли все статистические свойства при линейном преобразовании.

Линеаризуемые функции в 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	$y = b_0 \cdot e^{b_1 \cdot x}$	$\ln(y) = \ln(b_0) + b_1 \cdot x$

$\Longrightarrow$ Вычисление в SPSS (подгонка кривой)

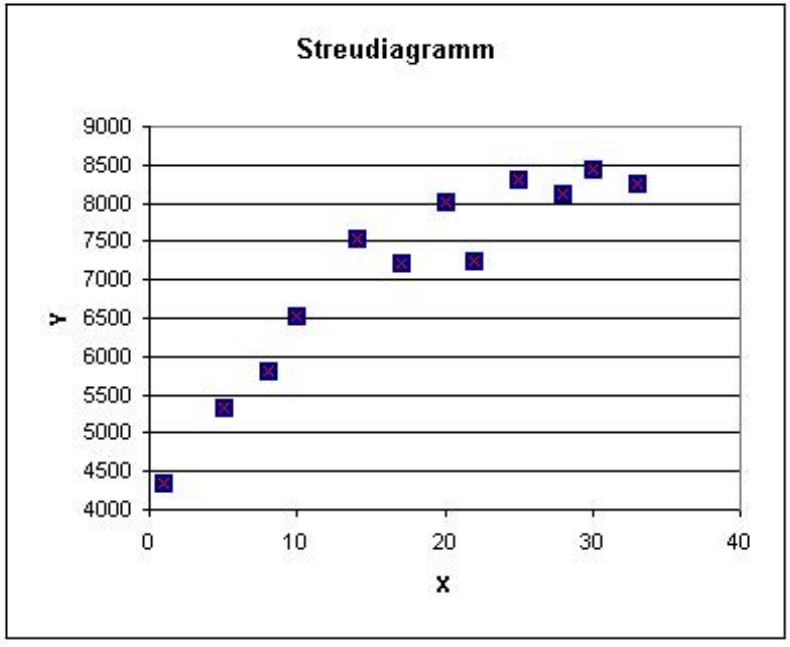
Ход кривых


Linear
Logarithmisch
Invers
Quadratisch
Kubisch
Exponent
1
2
3
3
Zusammengesetzt
1
2
3
S
Logistisch
Wachstum
Exponentiell

Пример

Пример: концентраты (50 кг/год) X и молочная продуктивность (л/год) Y за 12 лет

	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



Источник: von Auer, Эконометрика, Берлин, Гейдельберг, Нью-Йорк, 1999

Пример – подход

Шаги:

1. Выбор линеаризуемой функции по графику
   $\Rightarrow$ логарифмическая, экспоненциальная
2. Преобразование исходных значений
3. Решение линеаризованной задачи
4. Формулировка нелинейного уравнения оценки

Пример – линеаризация и решение линейной задачи

Логарифмическая: $y = b_0 + b_1 ln(x) + u$

Преобразование: $x \to ln(x) \Rightarrow z = ln(x)$

Линейные коэффициенты: $b_0 = 3818{,}334$, $b_1 = 1268{,}803$

Линейное уравнение регрессии: $y = 3818{,}334 + 1268{,}803\, ln(x) + u$

Нелинейное уравнение: $e^y = e^{(3818{,}334+u)} x^{1268{,}803}$

Пример – преобразование исходных значений

Пример: концентраты (50 кг/год) X и молочная продуктивность (л/год) Y за 12 лет

	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

$\Longrightarrow$ здесь нужно только $ln(y)$

Пример – линеаризация и решение (экспоненциальная)

Экспоненциальная: $y = e^{b_0 + b_1 x + u} \Rightarrow ln(y) = b_0 + b_1 x + u$
Преобразование: $y \to ln(y) \Rightarrow z = ln(y)$
Коэффициенты: $b_0 = 8{,}5236,\ b_1 = 0{,}018315$

Линейное уравнение: $ln(y) = 8{,}5236 + 0{,}018315\, x + u$

Нелинейное уравнение: $y = e^{8{,}5236 + 0{,}018315\, x + u}$

Содержание

- 3.3.1 Постановка задачи
- 3.3.2 Линеаризация
- **3.3.3 Эластичности**

Эластичности

Предельные изменения $\frac{dy}{dx}$:

Тип функции	**Уравнение**	**Предельные функции**
логарифмическая	$y = b_0 + b_1 ln(x) + u$	$b_1 x^{-1}$
экспоненциальная	$y = e^{b_0 + b_1 x + u}$	$b_1 e^{b_0 + b_1 x}$
обратная	$y = b_0 - b_1 \frac{1}{x} + u$	$-b_1 x^{-2}$
степенная	$y = b_0 x^{b_1} e^u$	$b_0 b_1 x^{b_1 - 1}$

Эластичности

Эластичности $\frac{dy}{dx} \cdot \frac{x}{y}$:

Тип	**Уравнение**	**Предельные**	**Эластичности**
лог.	$y = b_0 + b_1 ln(x)$	$b_1 x^{-1}$	$b_1 [b_0 + b_1 ln(x)]^{-1}$
эксп.	$y = e^{b_0 + b_1 x}$	$b_1 e^{b_0 + b_1 x}$	$b_1 x$
обр.	$y = b_0 - b_1 \frac{1}{x}$	$-b_1 x^{-1}$	$-b_1 [b_0 x + b_1]^{-1}$
степ.	$y = b_0 x^{b_1} e^u$	$b_0 b_1 x^{b_1 - 1}$	b_1

Введение в SPSS – темы

Статистические темы: дескриптивная статистика

- Подгонка кривой
- Нелинейная регрессия

Нелинейная регрессия – пример

с Excel:

Пример: ???

Задание: определить уравнение регрессии и меры качества

с SPSS:

Пример: ???

Задание: оценить коэффициенты, анализ ANOVA и других таблиц, графическое представление результатов

2. Корреляционный и регрессионный анализ – Содержание

- 3.1 Корреляционный анализ
- 3.2 Простая линейная регрессия
- 3.3 Простая нелинейная регрессия
- **3.4 Множественная регрессия**

Содержание

- **3.4.1 Определение и вывод**
- 3.4.2 Меры качества
- 3.4.3 Пример

Определение

Определение: Y и X_j, $j = 1, 2, ..., m$ – метрически масштабированные признаки. Значения получены у n единиц:

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

Линейная функция $y^* = b_0 + \sum_{j=1}^{m} b_j \cdot x_j$ для описания статистической зависимости Y от X_j называется **множественной линейной регрессией**.

X_j – регрессоры, независимые или экзогенные переменные
Y – регрессант, зависимая или эндогенная переменная
b_0 – константа регрессии (intercept)
b_j – линейные коэффициенты регрессии, $\mathbf{b} = (b_1, ..., b_m)$

$y^* = b_0 + \mathbf{bx}$

Определение

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

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

Вывод

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

Задача на экстремум:

$S(b) = u^T u = (y - Xb)^T (y - Xb) \to MIN$

...

Система нормальных уравнений:

$X^T Xb = X^T y$

$\Longrightarrow$

Решение:

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

Содержание

- 3.4.1 Определение и вывод
- **3.4.2 Меры качества**
- 3.4.3 Пример

Меры качества

Сумма квадратов остатков:

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

Коэффициент детерминации:

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

Скорректированный коэффициент детерминации:

$\overline{R^2} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$, где k – число экзогенных переменных

Содержание

- 3.4.1 Определение и вывод
- 3.4.2 Меры качества
- **3.4.3 Пример**

Пример

	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

Пример из лекции 3

Введение в SPSS – темы

Статистические темы: дескриптивная статистика

- Линейная регрессия
- Методы выбора независимых переменных

Введение в SPSS – темы

1. Включение (регрессия). Процедура выбора переменных, при которой все переменные блока включаются за один шаг.
2. Пошаговый отбор. На каждом шаге включается независимая переменная, ещё не входящая в уравнение, если критерий отбора выполнен.
3. Исключение (выбор переменных). Метод выбора переменных, при котором все переменные блока исключаются за один шаг.
4. Обратное исключение. Метод выбора переменных, при котором все переменные включаются в уравнение, а затем последовательно исключаются.
5. Прямой отбор. Метод пошагового выбора переменных, при котором переменные последовательно включаются в модель.

Множественная линейная регрессия – пример

с Excel:

Пример 2: Zinssaetze 2007_2.xls

Задание: определить уравнение регрессии и меры качества

с SPSS:

Пример 2: Zinssaetze 2007_2.xls

Задание: оценить коэффициенты, анализ ANOVA и других таблиц, графическое представление результатов

Множественная нелинейная регрессия – пример

с Excel:

Пример 2: Производственная функция Кобба-Дугласа

Задание: определить уравнение регрессии и меры качества, экономическая интерпретация

с SPSS:

Пример 2: Производственная функция Кобба-Дугласа

Задание: оценить коэффициенты, анализ ANOVA и других таблиц, графическое представление, экономическая интерпретация