Karlsruher Institut für Technologie

Institut für Volkswirtschaftslehre (ECON)

Lehrstuhl für Ökonometrie und Statistik

Kapitel X - Lineare Regression

Deskriptive Statistik

Prof. Dr. W.-D. Heller
Hartwig Senska
Carlo Siebenschuh

Agenda

1. **Untersuchung auf lineare Abhängigkeit**
2. Methode der kleinsten Quadrate

Untersuchung der Abhängigkeit

Hat die Untersuchung zweier Merkmale auf einer statistischen Masse ergeben, dass eine Abhängigkeit zwischen diesen Merkmalen besteht, so stellt sich unmittelbar die Frage, ob das Datenmaterial Aussagen über die Art und die Stärke der Abhängigkeit zulässt.

Sind beide Merkmale quantitativ, so bestehen die Beobachtungen aus Paaren reeller Zahlen. Einen intuitiven Eindruck über die Art der Abhängigkeit erhält man durch das **Streuungsdiagramm**.

Untersuchung der Abhängigkeit

Hat die Untersuchung zweier Merkmale auf einer statistischen Masse ergeben, dass eine Abhängigkeit zwischen diesen Merkmalen besteht, so stellt sich unmittelbar die Frage, ob das Datenmaterial Aussagen über die Art und die Stärke der Abhängigkeit zulässt.

Sind beide Merkmale quantitativ, so bestehen die Beobachtungen aus Paaren reeller Zahlen. Einen intuitiven Eindruck über die Art der Abhängigkeit erhält man durch das **Streuungsdiagramm**.

Untersuchung der Abhängigkeit

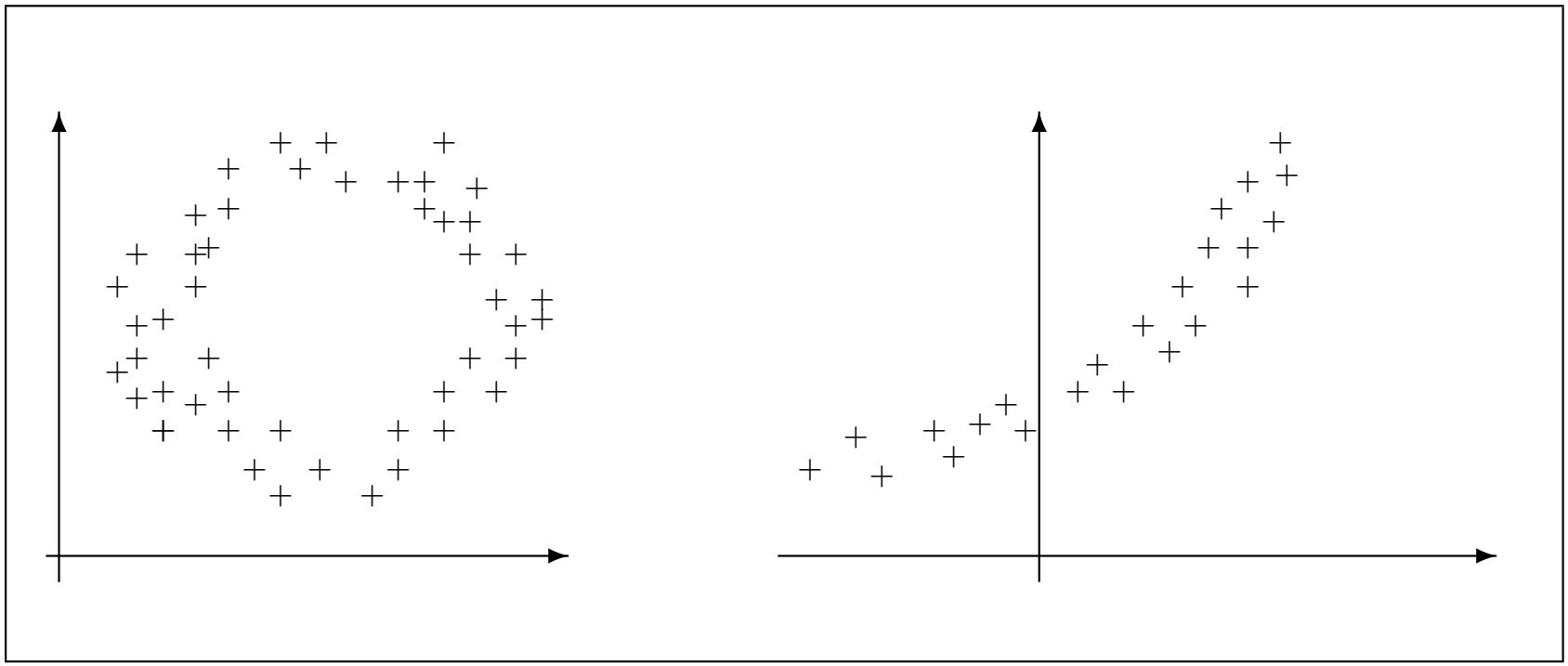


Abbildung 10.1 - Streuungsdiagramme mit Ähnlichkeit zu einer geometrischen Figur bzw. einem Funktionsgraph.

Es drängt sich die Überlegung auf, dass die Daten sich dadurch ergeben, dass ein funktionaler Zusammenhang zwischen den beiden Merkmalen vorliegt, der z.B. durch Messfehler, individuelle Besonderheiten oder Abhängigkeiten mit anderen Merkmalen überlagert ist.

Untersuchung der Abhängigkeit

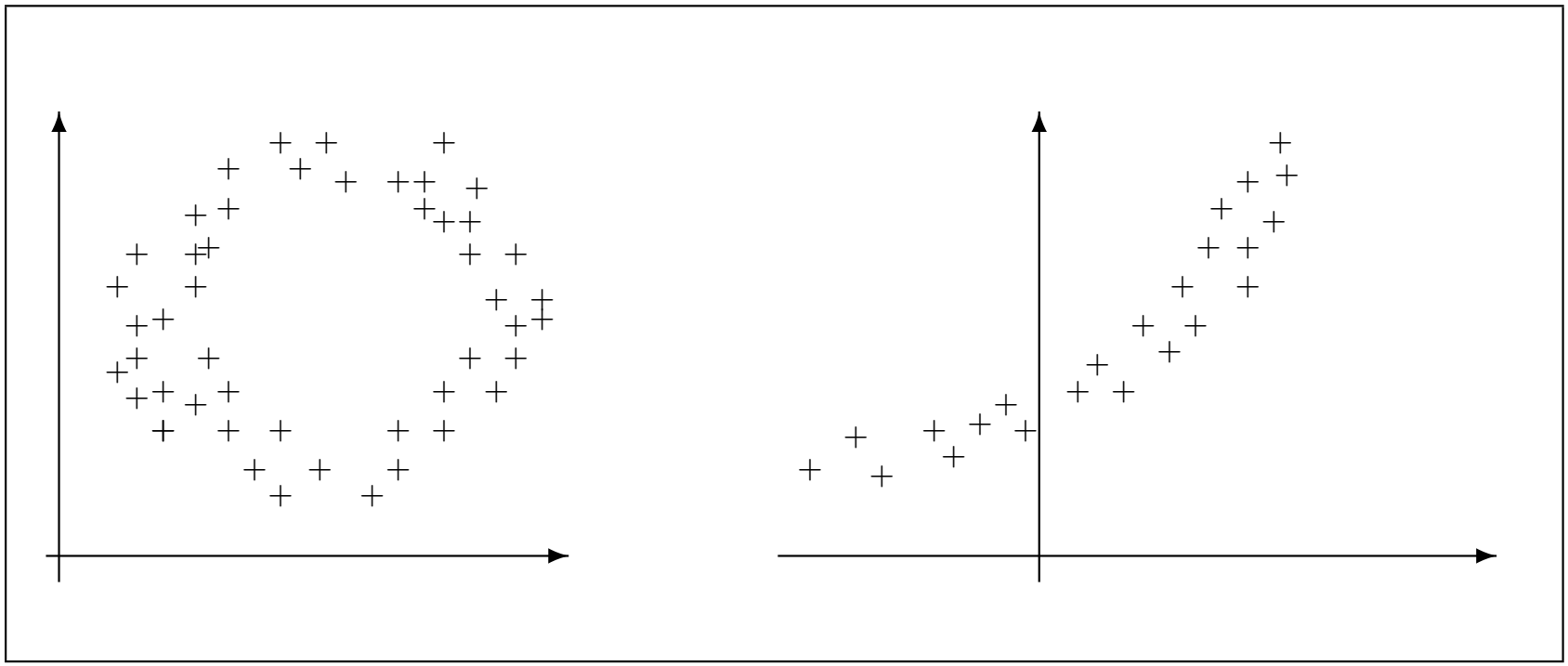


Abbildung 10.1 - Streuungsdiagramme mit Ähnlichkeit zu einer geometrischen Figur bzw. einem Funktionsgraph.

Es drängt sich die Überlegung auf, dass die Daten sich dadurch ergeben, dass ein funktionaler Zusammenhang zwischen den beiden Merkmalen vorliegt, der z.B. durch Messfehler, individuelle Besonderheiten oder Abhängigkeiten mit anderen Merkmalen überlagert ist.

Untersuchung der Abhängigkeit

Annahmne
Den Beobachtungspaaren $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ liegt ein linearer Zusammenhang ("**Trend**") zugrunde.

Dieser lässt sich ausdrücken durch eine Funktion vom Typ

$$y = mx + b,$$

wobei die Parameter m und b den Anstieg und den y-Achsen-Abschnitt angeben.

Aufgabe ist also, mit dem Datenmaterial die unbekannten Werte m und b - zumindest näherungsweise - zu bestimmen.

Untersuchung der Abhängigkeit

Annahmne
Den Beobachtungspaaren $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ liegt ein linearer Zusammenhang ("**Trend**") zugrunde.

Dieser lässt sich ausdrücken durch eine Funktion vom Typ

$$y = mx + b,$$

wobei die Parameter m und b den Anstieg und den y-Achsen-Abschnitt angeben.

Aufgabe ist also, mit dem Datenmaterial die unbekannten Werte m und b - zumindest näherungsweise - zu bestimmen.

Untersuchung der Abhängigkeit

Annahmne
Den Beobachtungspaaren $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ liegt ein linearer Zusammenhang ("**Trend**") zugrunde.

Dieser lässt sich ausdrücken durch eine Funktion vom Typ

$$y = mx + b,$$

wobei die Parameter m und b den Anstieg und den y-Achsen-Abschnitt angeben.

Aufgabe ist also, mit dem Datenmaterial die unbekannten Werte m und b - zumindest näherungsweise - zu bestimmen.

Agenda

1. Untersuchung auf lineare Abhängigkeit
2. **Methode der kleinsten Quadrate**

Methode der kleinsten Quadrate

Angenommen, m und b wären bekannt, so ließe sich zu jedem x-Wert der "Trendwert" $\hat{y}_i$ ermitteln:

$$\hat{y}_i = mx_i + b.$$

Daraus ergäbe sich dann auch die Störgröße

$$y_i - \hat{y}_i = y_i - mx_i - b.$$

Das Prinzip der **linearen Regression** - auch **Methode der kleinsten Quadrate** genannt - ist:
Bestimme m und b so, dass die Summe der quadrierten Störgrößen minimiert wird:

$$\min \sum_{i=1}^{n} (y_i - mx_i - b)^2$$

Methode der kleinsten Quadrate

Angenommen, m und b wären bekannt, so ließe sich zu jedem x-Wert der "Trendwert" $\hat{y}_i$ ermitteln:

$$\hat{y}_i = mx_i + b.$$

Daraus ergäbe sich dann auch die Störgröße

$$y_i - \hat{y}_i = y_i - mx_i - b.$$

Das Prinzip der **linearen Regression** - auch **Methode der kleinsten Quadrate** genannt - ist:
Bestimme m und b so, dass die Summe der quadrierten Störgrößen minimiert wird:

$$\min \sum_{i=1}^{n} (y_i - mx_i - b)^2$$

Methode der kleinsten Quadrate

Angenommen, m und b wären bekannt, so ließe sich zu jedem x-Wert der "Trendwert" $\hat{y}_i$ ermitteln:

$$\hat{y}_i = mx_i + b.$$

Daraus ergäbe sich dann auch die Störgröße

$$y_i - \hat{y}_i = y_i - mx_i - b.$$

Das Prinzip der **linearen Regression** - auch **Methode der kleinsten Quadrate** genannt - ist:
Bestimme m und b so, dass die Summe der quadrierten Störgrößen minimiert wird:

$$\min \sum_{i=1}^{n} (y_i - mx_i - b)^2$$

Methode der kleinsten Quadrate

Nach partieller Differentiation nach m und b ergibt sich aus der notwendigen Bedingung:

$$\hat{m} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \bar{y} - \hat{m}\bar{x}$$

(hinreichende Bedingung prüfen !)

Methode der kleinsten Quadrate

Beispiel 10.1
Gegeben sei die folgende Liste von Beobachtungspaaren:

$(7;8), (7;7), (8;9), (10;11), (11;12), (14;15), (17;18), (17;16), (19;20), (18;19)$.

Dann ist:

i											$\sum$
x_i	7	7	8	10	11	14	17	17	19	18	128
x_i^2	49	49	64	100	121	196	289	289	361	324	1842
y_i	8	7	9	11	12	15	18	16	20	19	135
$x_i y_i$	56	49	72	110	132	210	306	272	380	342	1929

und damit

$$\hat{m} = \frac{10 \cdot 1929 - 128 \cdot 135}{10 \cdot 1842 - 128^2} = \frac{2010}{2036} = 0.987,$$

$$\hat{b} = \frac{1842 \cdot 135 - 128 \cdot 1929}{2036} = \frac{1758}{2036} = 0.863.$$

Methode der kleinsten Quadrate

Beispiel 10.1
Gegeben sei die folgende Liste von Beobachtungspaaren:

$(7;8), (7;7), (8;9), (10;11), (11;12), (14;15), (17;18), (17;16), (19;20), (18;19).$

Dann ist:

i											$\sum$
x_i	7	7	8	10	11	14	17	17	19	18	128
x_i^2	49	49	64	100	121	196	289	289	361	324	1842
y_i	8	7	9	11	12	15	18	16	20	19	135
$x_i y_i$	56	49	72	110	132	210	306	272	380	342	1929

und damit

$$\hat{m} = \frac{10 \cdot 1929 - 128 \cdot 135}{10 \cdot 1842 - 128^2} = \frac{2010}{2036} = 0.987,$$

$$\hat{b} = \frac{1842 \cdot 135 - 128 \cdot 1929}{2036} = \frac{1758}{2036} = 0.863.$$

Methode der kleinsten Quadrate

Beispiel 10.1
Gegeben sei die folgende Liste von Beobachtungspaaren:

$(7;8), (7;7), (8;9), (10;11), (11;12), (14;15), (17;18), (17;16), (19;20), (18;19).$

Dann ist:

i											$\sum$
x_i	7	7	8	10	11	14	17	17	19	18	128
x_i^2	49	49	64	100	121	196	289	289	361	324	1842
y_i	8	7	9	11	12	15	18	16	20	19	135
$x_i y_i$	56	49	72	110	132	210	306	272	380	342	1929

und damit

$$\hat{m} = \frac{10 \cdot 1929 - 128 \cdot 135}{10 \cdot 1842 - 128^2} = \frac{2010}{2036} = 0.987,$$

$$\hat{b} = \frac{1842 \cdot 135 - 128 \cdot 1929}{2036} = \frac{1758}{2036} = 0.863.$$

Methode der kleinsten Quadrate

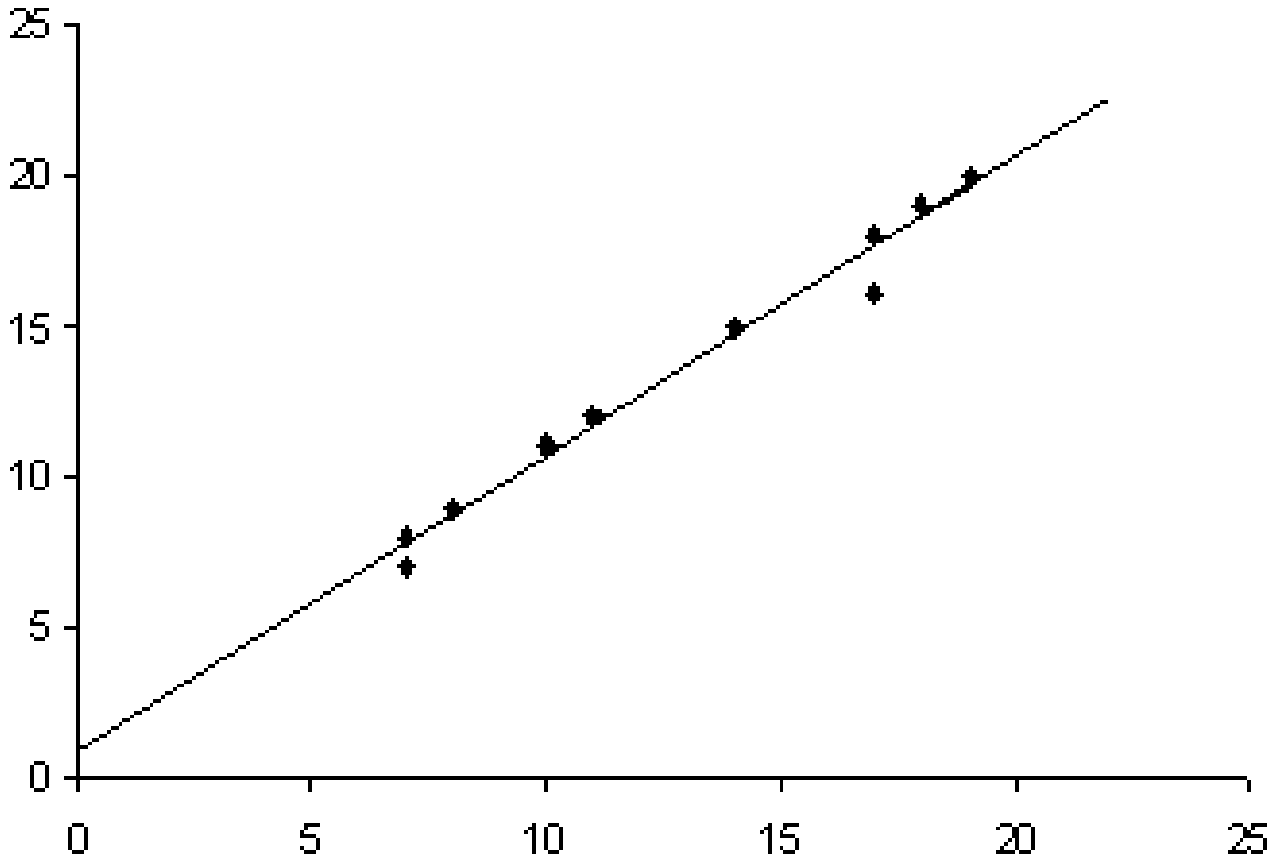


Abbildung 10.2 - Streuungsdiagramm und Regressionsgerade zu Beispiel 10.1

Methode der kleinsten Quadrate

Bemerkung: Die Methode der kleinsten Quadrate lässt sich auch bei nichtlinearen Zusammenhängen anwenden. Hierbei wird zunächst, soweit möglich, eine Transformation der Daten vorgenommen, so dass die transformierten Daten zu einem linearen Zusammenhang führen. Für die transformierten Daten lassen sich die Parameter dann mit linearer Regression bestimmen. Anschließend wird die Transformation wieder rückgängig gemacht.

Methode der kleinsten Quadrate

Beispiel 10.2

Gegeben sind die Beobachtungspaare

$$(0; 1.1), (1; 2.5), (2; 8), (3; 25), (4; 65).$$

Aufgrund des starken Ansteigens der y-Werte wird eine exponentielle Abhängigkeit vermutet:

$$y = Ce^{\alpha x}$$

mit unbekanntem C und α. Durch Logarithmieren erhält man:

$$\ln(y) = \ln(Ce^{\alpha x}) = \ln C + \alpha x, \ (\text{linear})$$

Methode der kleinsten Quadrate

Beispiel 10.2
Gegeben sind die Beobachtungspaare

$$(0; 1.1), (1; 2.5), (2; 8), (3; 25), (4; 65).$$

Aufgrund des starken Ansteigens der y-Werte wird eine exponentielle Abhängigkeit vermutet:

$$y = Ce^{\alpha x}$$

mit unbekanntem C und α. Durch Logarithmieren erhält man:

$$\ln(y) = \ln(Ce^{\alpha x}) = \ln C + \alpha x, \ (\text{linear})$$

Methode der kleinsten Quadrate

Beispiel 10.2
Gegeben sind die Beobachtungspaare

$$(0; 1.1), (1; 2.5), (2; 8), (3; 25), (4; 65).$$

Aufgrund des starken Ansteigens der y-Werte wird eine exponentielle Abhängigkeit vermutet:

$$y = Ce^{\alpha x}$$

mit unbekanntem C und α. Durch Logarithmieren erhält man:

$$\ln(y) = \ln(Ce^{\alpha x}) = \ln C + \alpha x, \ (\text{linear})$$

Methode der kleinsten Quadrate

Beispiel 10.2

Die lineare Regression liefert dann:

i	1	2	3	4	5	$\sum$
x_i	0	1	2	3	4	10
x_i^2	0	1	4	9	16	30
y_i	1.1	2.5	8	25	65	
$\ln y_i$	.0953	.9163	2.0794	3.2189	4.1744	10.484
$x_i \ln y_i$	0	.9163	4.1589	9.6566	16.6975	31.429

$$\hat{m} = \hat{\alpha} = \frac{5 \cdot 31.429 - 10 \cdot 10.484}{5 \cdot 30 - 100} = \frac{52.304}{50} = 1.05,$$
$$\hat{b} = \widehat{\ln C} = \frac{30 \cdot 10.484 - 10 \cdot 31.429}{50} = 0.005, \ \hat{C} = 1.005.$$

Daraus ergeben sich die $\hat{y}_i$-Werte:

i	1	2	3	4	5
$\hat{y}_i$	1.005	2.872	8.207	23.453	67.020

Methode der kleinsten Quadrate

Beispiel 10.2

Die lineare Regression liefert dann:

i	1	2	3	4	5	$\sum$
x_i	0	1	2	3	4	10
x_i^2	0	1	4	9	16	30
y_i	1.1	2.5	8	25	65	
$\ln y_i$	.0953	.9163	2.0794	3.2189	4.1744	10.484
$x_i \ln y_i$	0	.9163	4.1589	9.6566	16.6975	31.429

$$\hat{m} = \hat{\alpha} = \frac{5 \cdot 31.429 - 10 \cdot 10.484}{5 \cdot 30 - 100} = \frac{52.304}{50} = 1.05,$$
$$\hat{b} = \widehat{\ln C} = \frac{30 \cdot 10.484 - 10 \cdot 31.429}{50} = 0.005, \ \hat{C} = 1.005.$$

Daraus ergeben sich die $\hat{y}_i$-Werte:

i	1	2	3	4	5
$\hat{y}_i$	1.005	2.872	8.207	23.453	67.020

Methode der kleinsten Quadrate

Beispiel 10.2

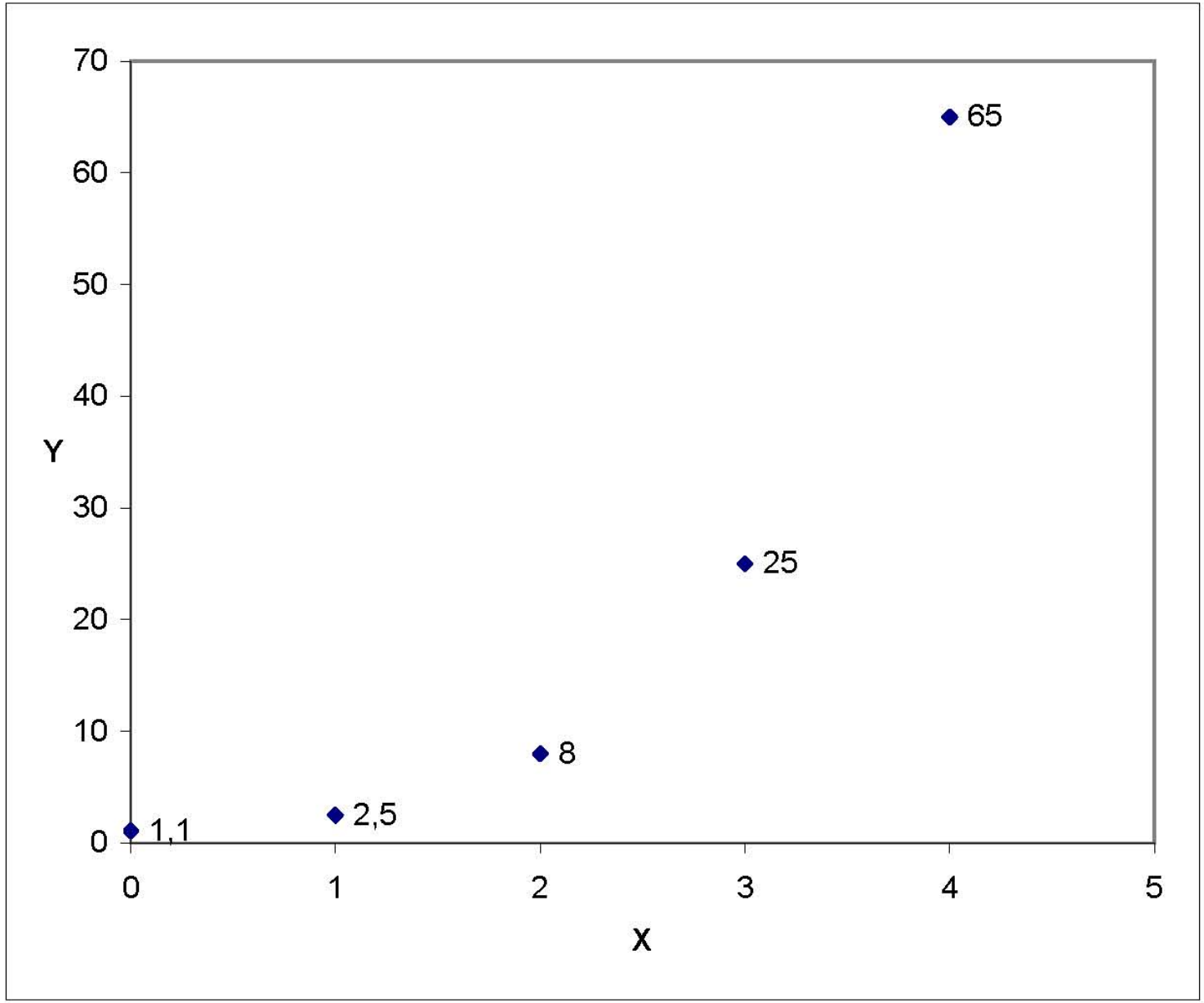
Die lineare Regression liefert dann:

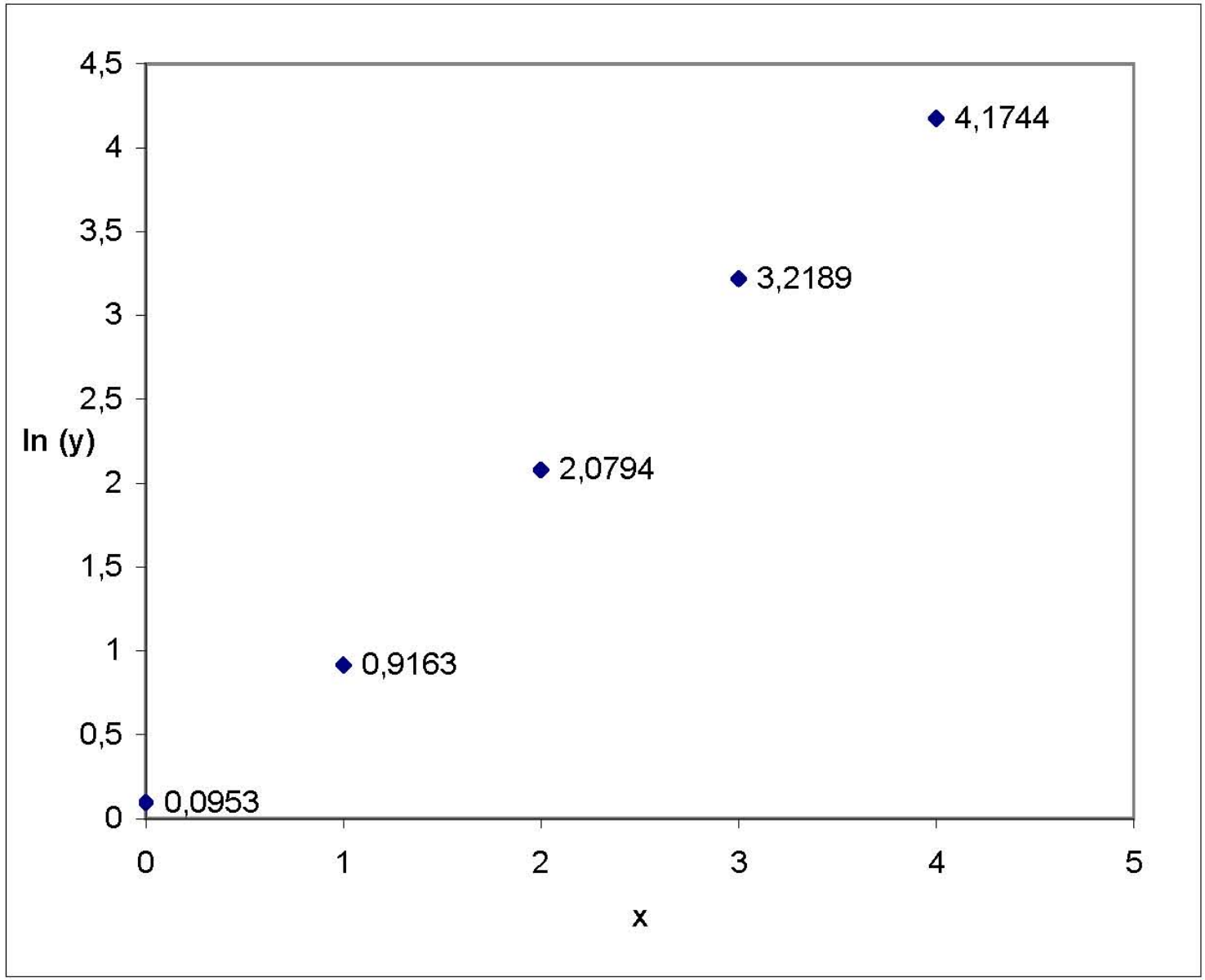
i	1	2	3	4	5	$\sum$
x_i	0	1	2	3	4	10
x_i^2	0	1	4	9	16	30
y_i	1.1	2.5	8	25	65	
$\ln y_i$	.0953	.9163	2.0794	3.2189	4.1744	10.484
$x_i \ln y_i$	0	.9163	4.1589	9.6566	16.6975	31.429

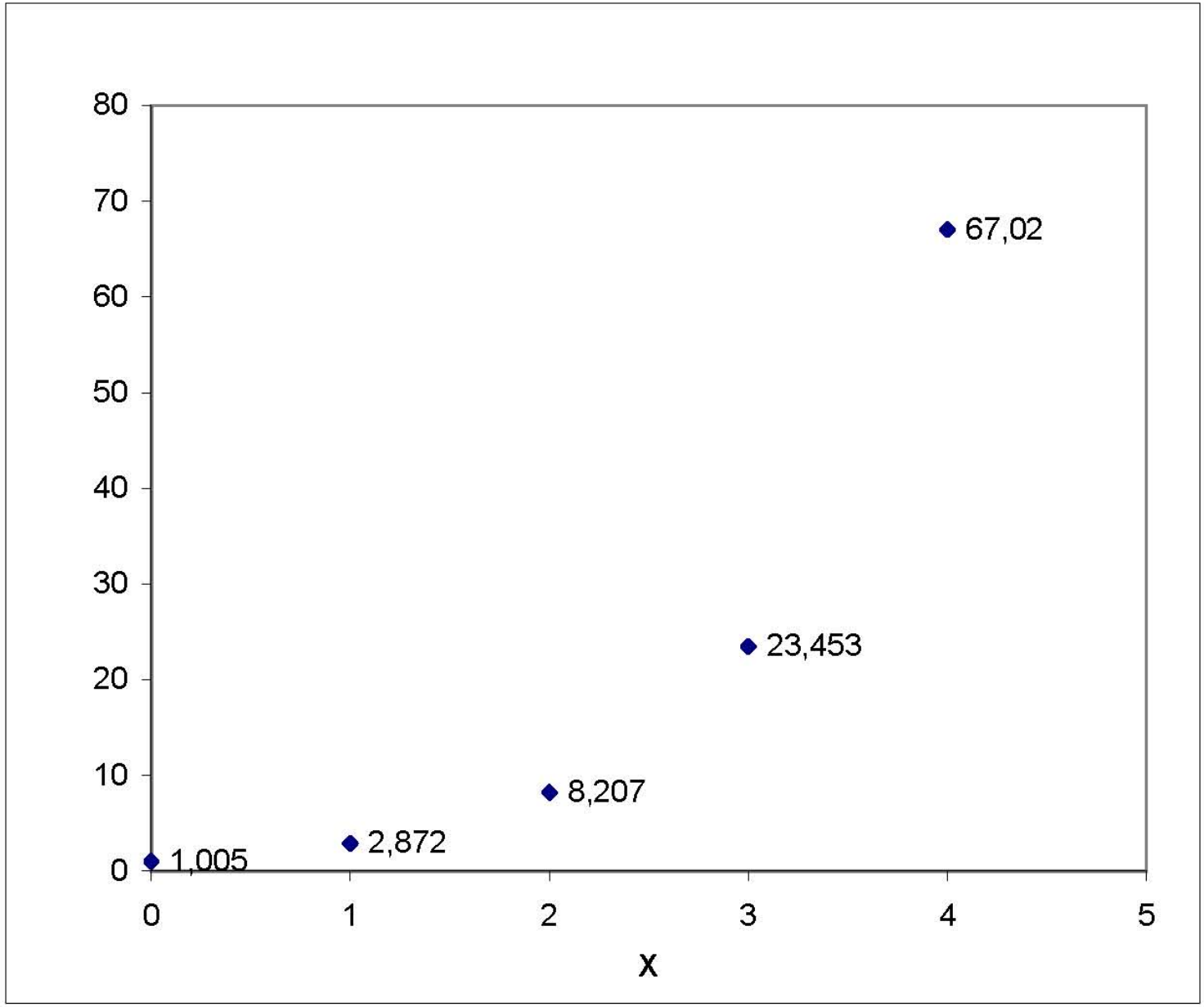
$$\hat{m} = \hat{\alpha} = \frac{5 \cdot 31.429 - 10 \cdot 10.484}{5 \cdot 30 - 100} = \frac{52.304}{50} = 1.05,$$
$$\hat{b} = \widehat{\ln C} = \frac{30 \cdot 10.484 - 10 \cdot 31.429}{50} = 0.005, \ \hat{C} = 1.005.$$

Daraus ergeben sich die $\hat{y}_i$-Werte:

i	1	2	3	4	5
$\hat{y}_i$	1.005	2.872	8.207	23.453	67.020


70
60
50
40
30
20
10
0
Y
65
25
8
2,5
1,1
0
1
2
3
4
5
X


4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
ln (y)
0
1
2
3
4
5
x
4,1744
3,2189
2,0794
0,9163
0,0953


80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
X
67,02
23,453
8,207
2,872
1,005