Regression

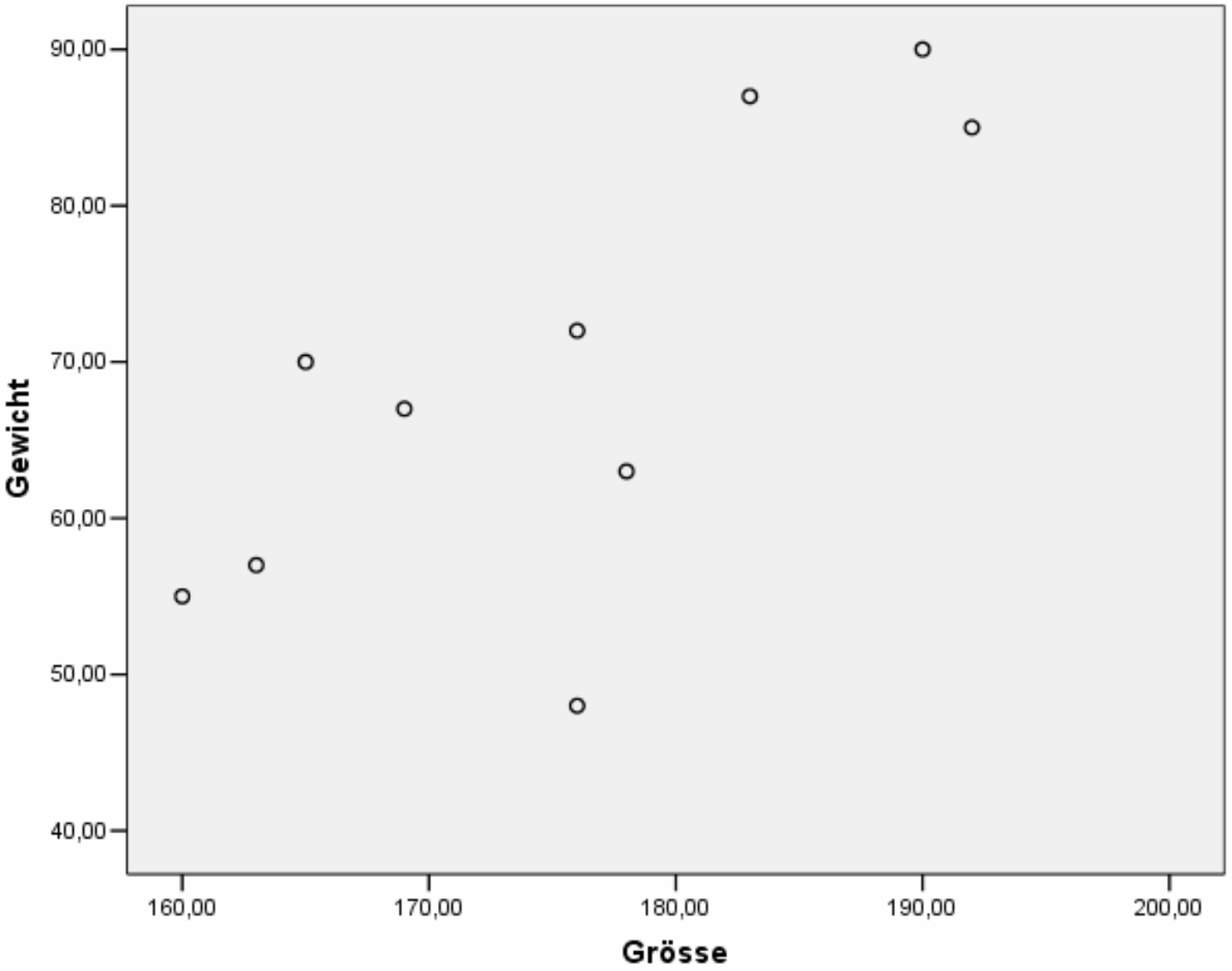
- PRE Konzept
- Regressionsgerade
- Determinationskoeffzient r^2
- Verhältnis von r und r^2
- Prognosen

Das Beispiel

Befragter	Körpergröße in cm	Körpergewicht in kg
A	160	55
B	163	57
C	165	70
D	169	67
E	176	48
F	176	72
G	178	63
H	183	87
I	190	90
J	192	85

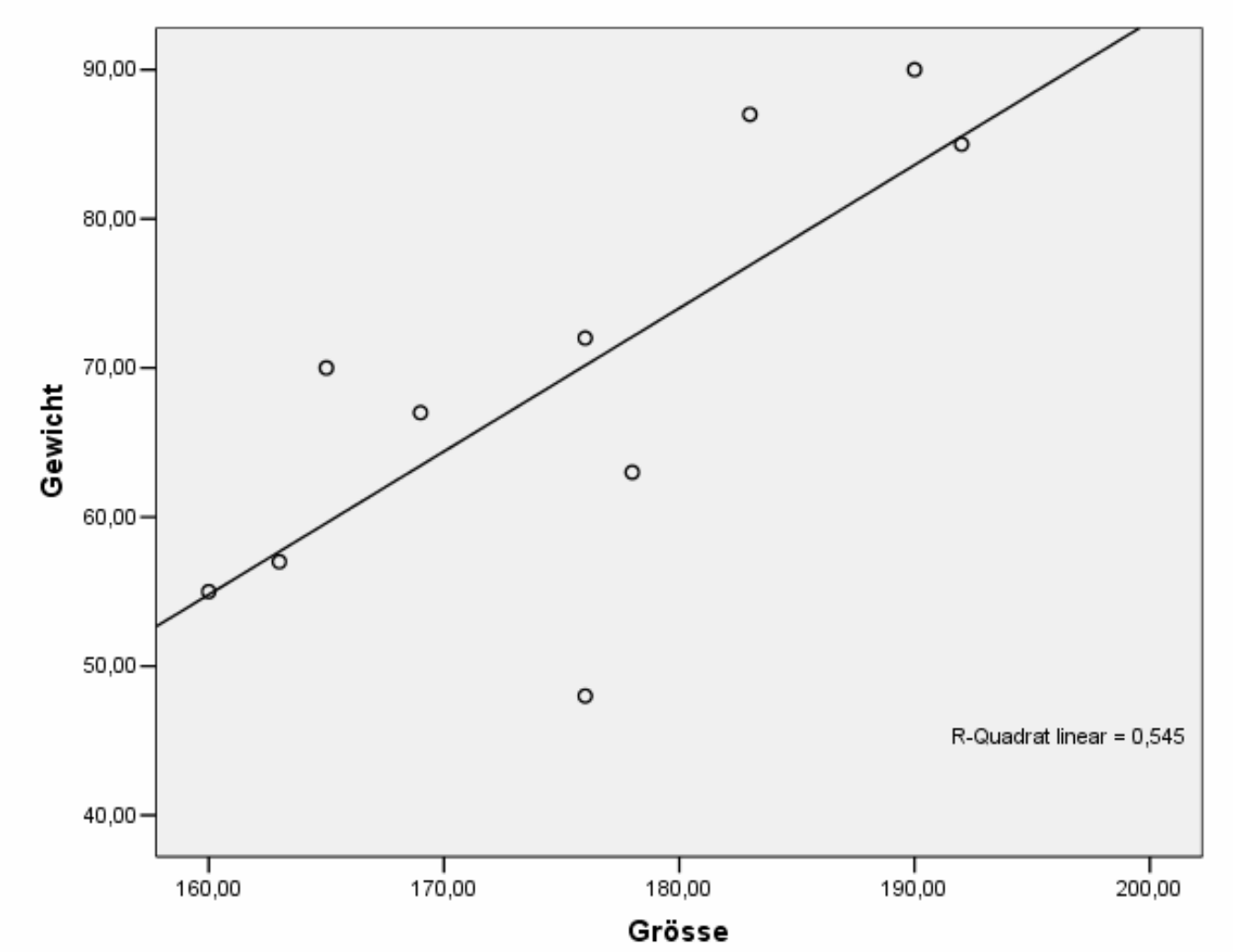
Voraussetzungen der Regression: metrische Variablen
linearer Zusammenhang

Streudiagramm mit Körpergröße und Körpergewicht



Die Punkte liegen nahe an einer Geraden.

Dies ist ein Zeichen für die Linearität der Beziehung.



Modell der proportionalen Fehlerreduktion (PRE Konzept)

Vorhersageregeln:

1. Eine Regel zur Vorhersage der abhängigen Variablen ohne Auswertung der Information über die unabhängige Variable

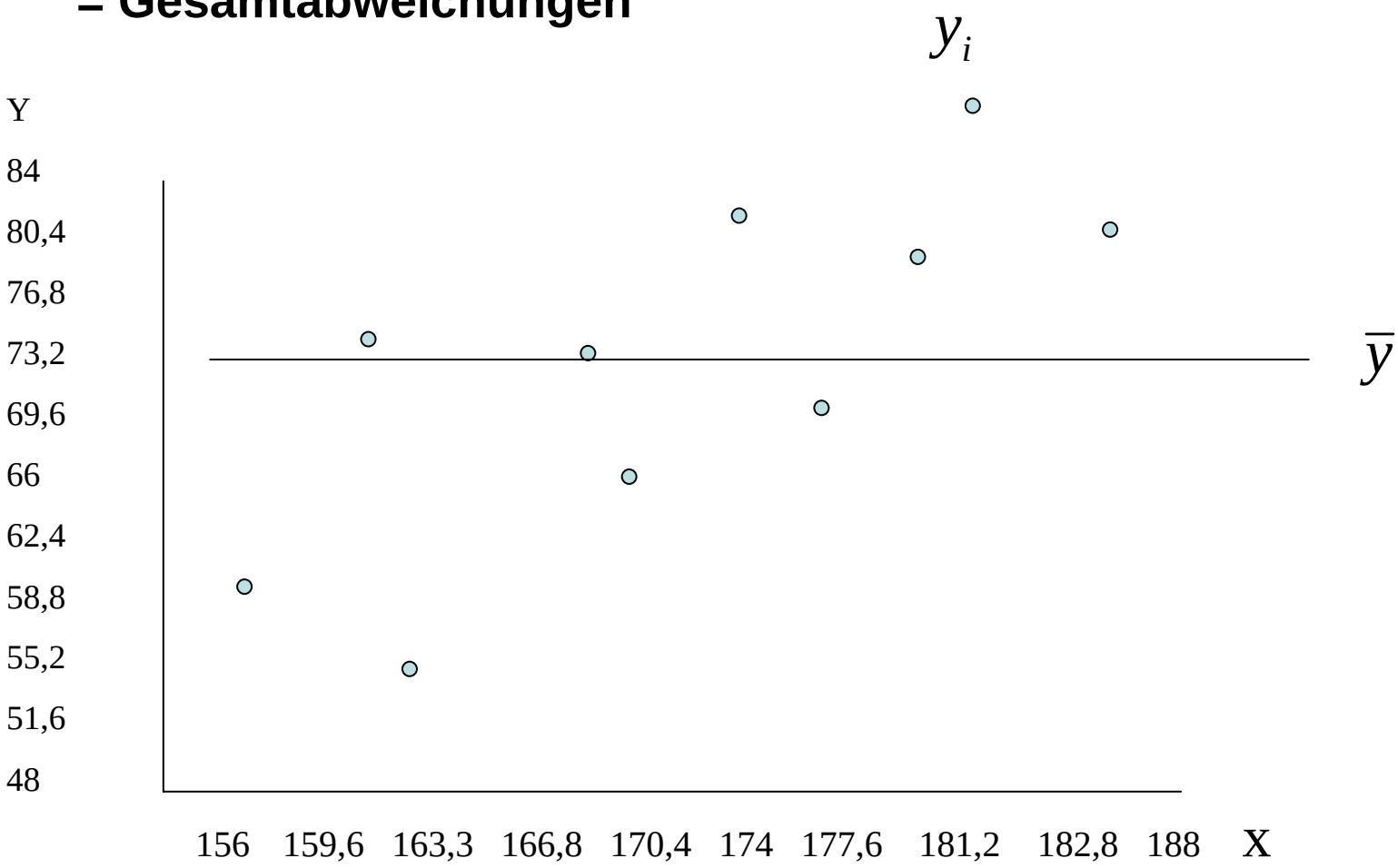
2. Eine Regel zur Vorhersage der abhängigen Variablen mit Auswertung der Information über die unabhängige Variable

$$\text{PRE - Maß} = \frac{E_1 - E_2}{E_1}$$

Der Determinationskoeffizient r^2 basiert auf dem PRE Konzept

Streudiagramm mit den Abweichungen vom Mittelwert

= Gesamtabweichungen



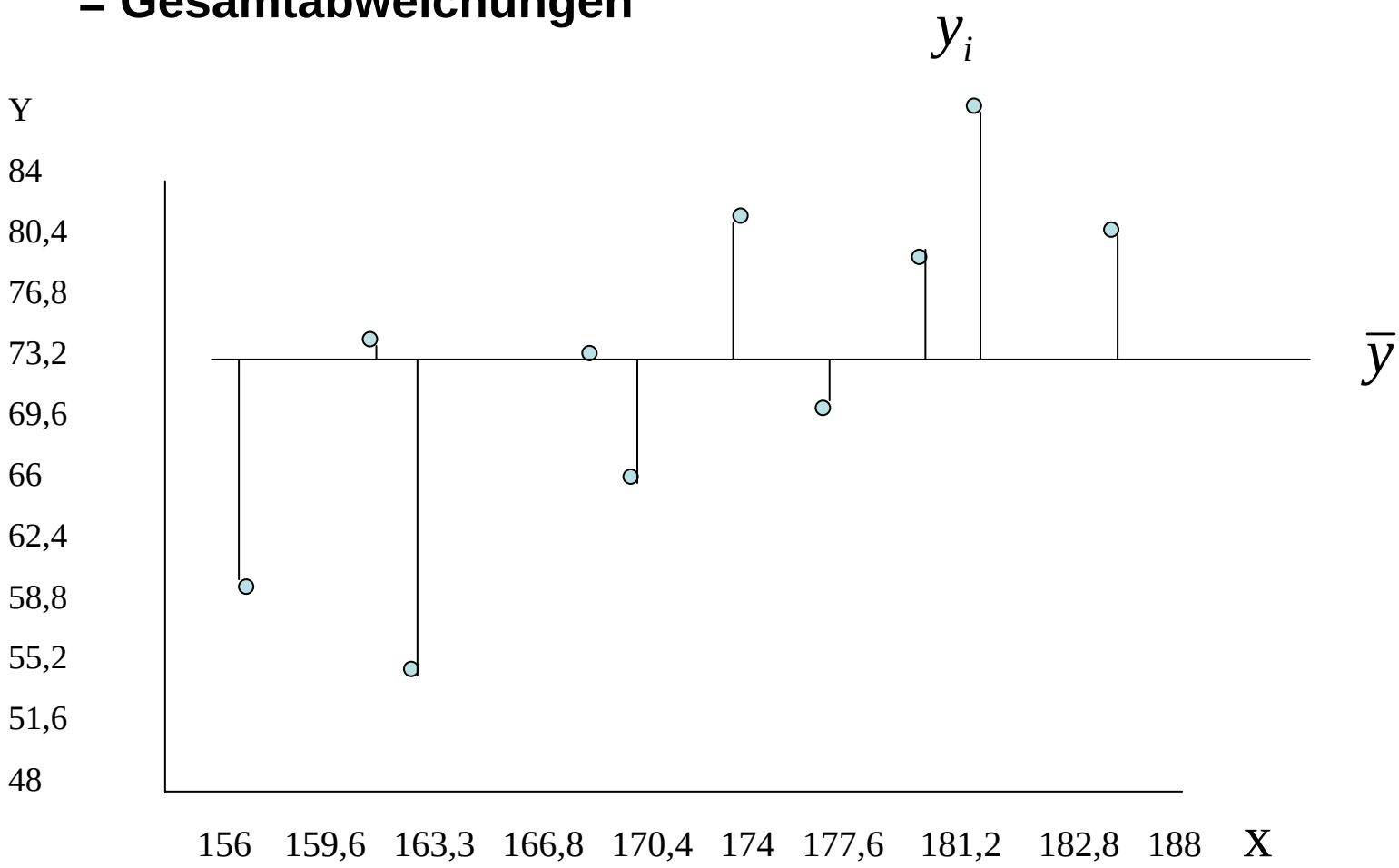
Fehler erster Art

Körpergewicht in kg y_i	$y_i - \bar{y}$
55	-14,4
57	-12,4
70	0,6
67	-2,4
48	-21,4
72	2,6
63	-6,4
87	17,6
90	20,6
85	15,6
$\Sigma = \mathbf{694}$	$\Sigma = \mathbf{0}$

$$\bar{y} = \frac{694}{10} = 69{,}4$$

Streudiagramm mit den Abweichungen vom Mittelwert

= Gesamtabweichungen



Logik des Determinationskoeffizienten

1. Vorhersage der abhängigen Variablen auf der Basis ihrer eigenen Verteilung: Sage mir für jede Untersuchungseinheit das arithmetische Mittel vorher

Die Fehlerdefinition: Bei der Vorhersage nach Regel 1 ist der Vorhersagefehler die Summe der quadrierten Abweichungen der y-Werte von ihrem arithmetischen Mittel

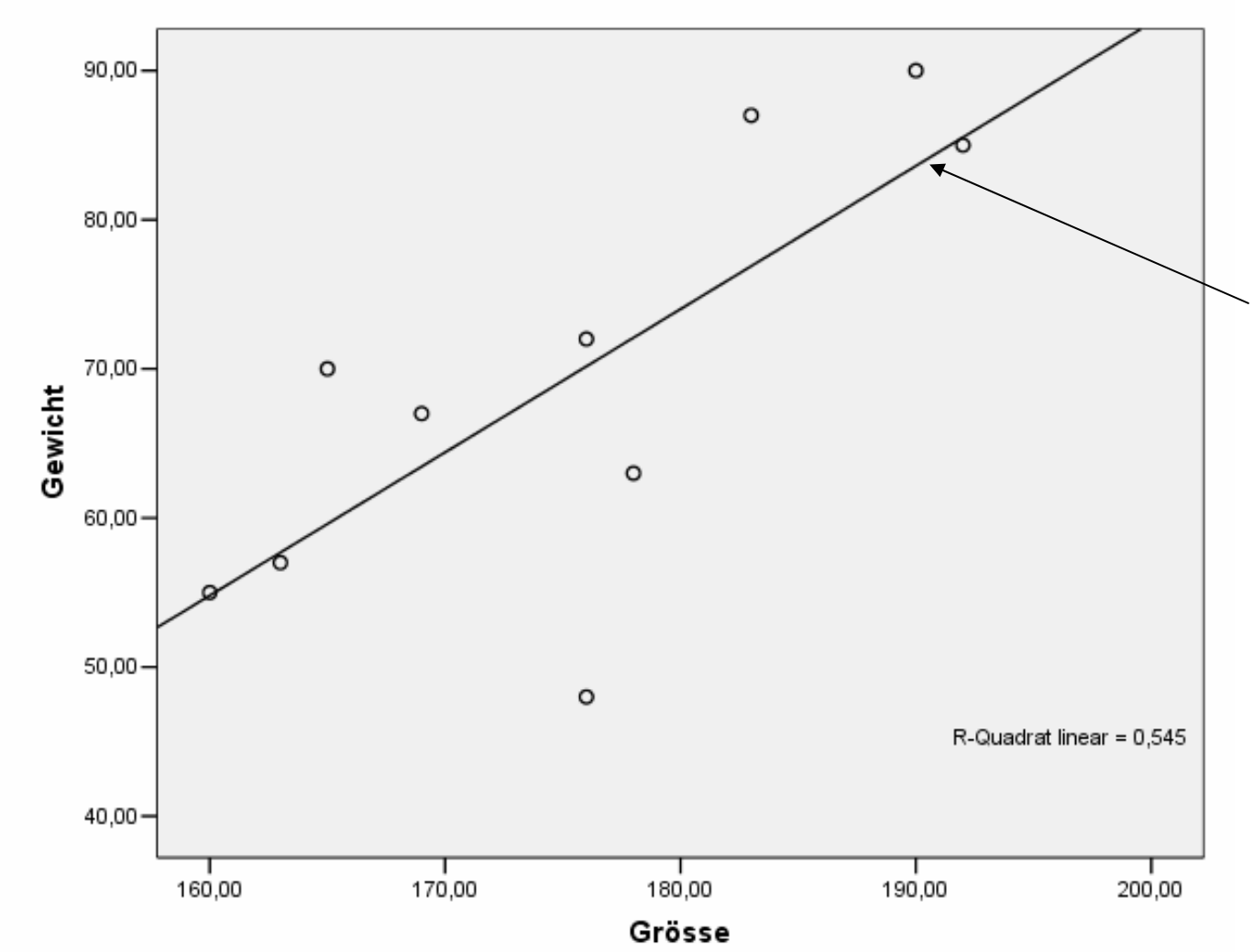
$$E_1 = \Sigma (y_i - \bar{y})^2$$

2. Vorhersage der abhängigen Variablen auf der Basis der unabhängigen Variable: Sage mir für jede Untersuchungseinheit den Regressionswert vorher

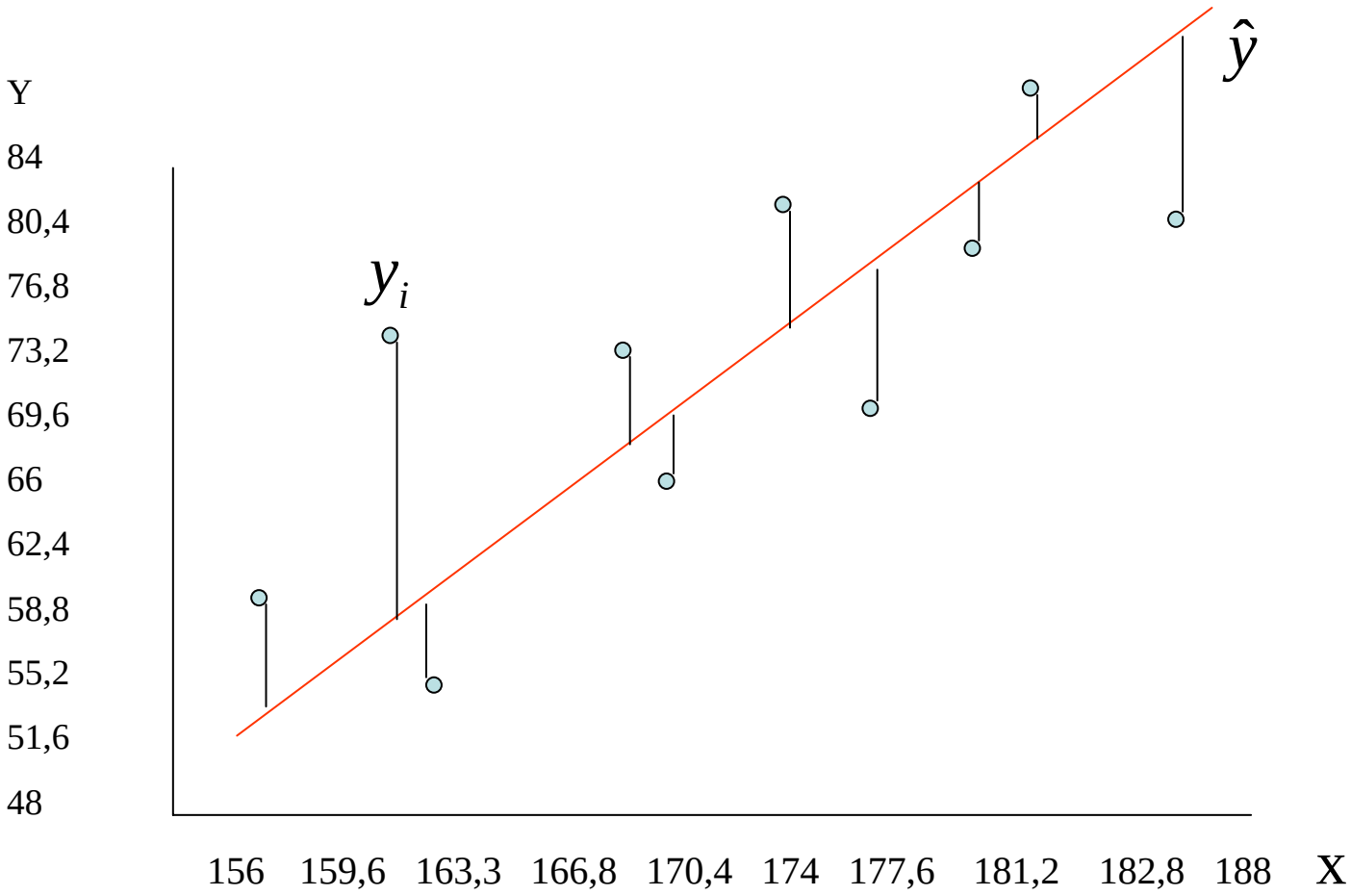
Die Fehlerdefinition: Bei der Vorhersage nach Regel 2 ist der Vorhersagefehler die Summe der quadrierten Abweichungen der y-Werte von der Regressionsgeraden

$$E_2 = \Sigma (y_i - y'_i)^2$$

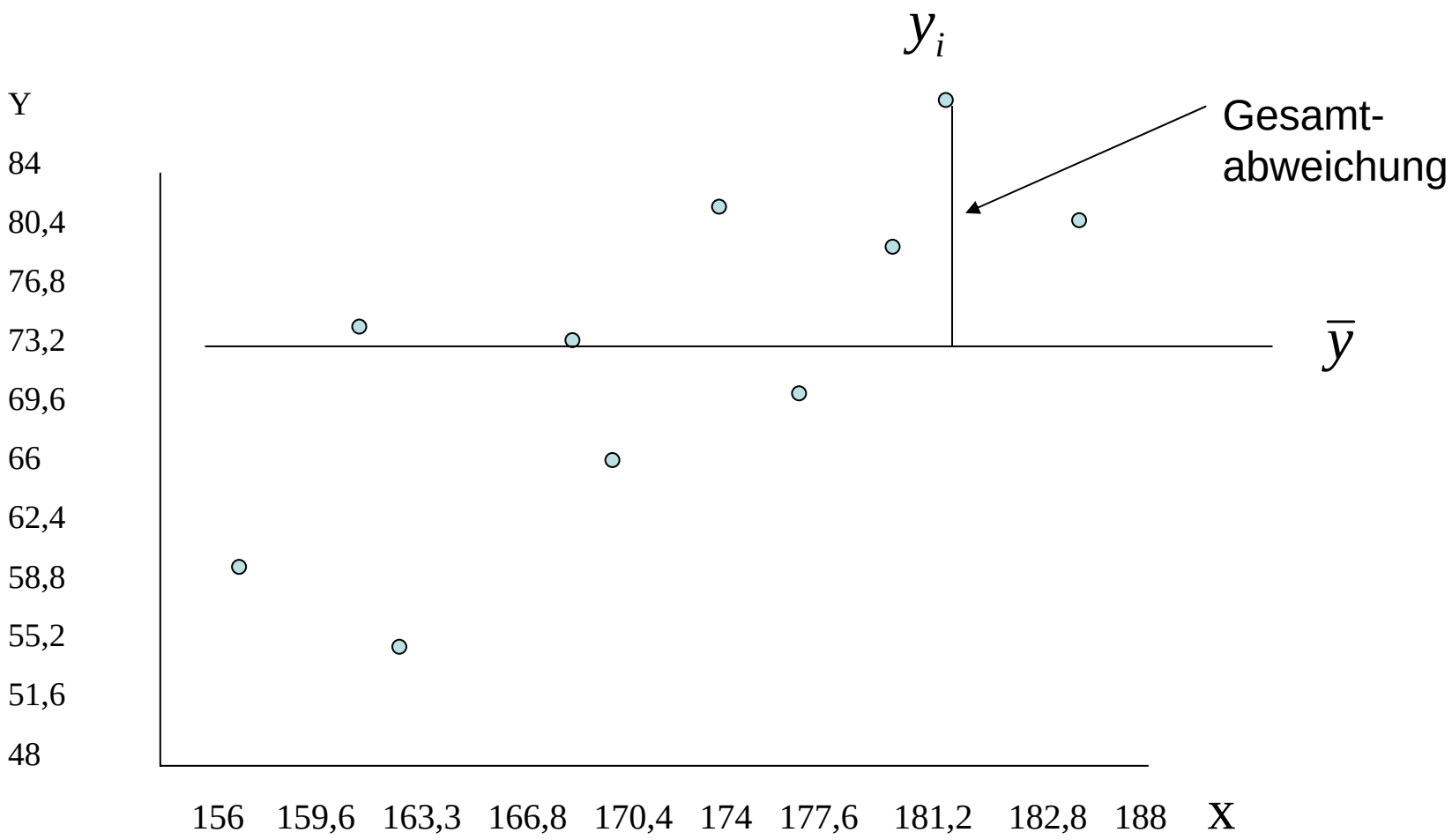
Streudiagramm mit Regressionslinie



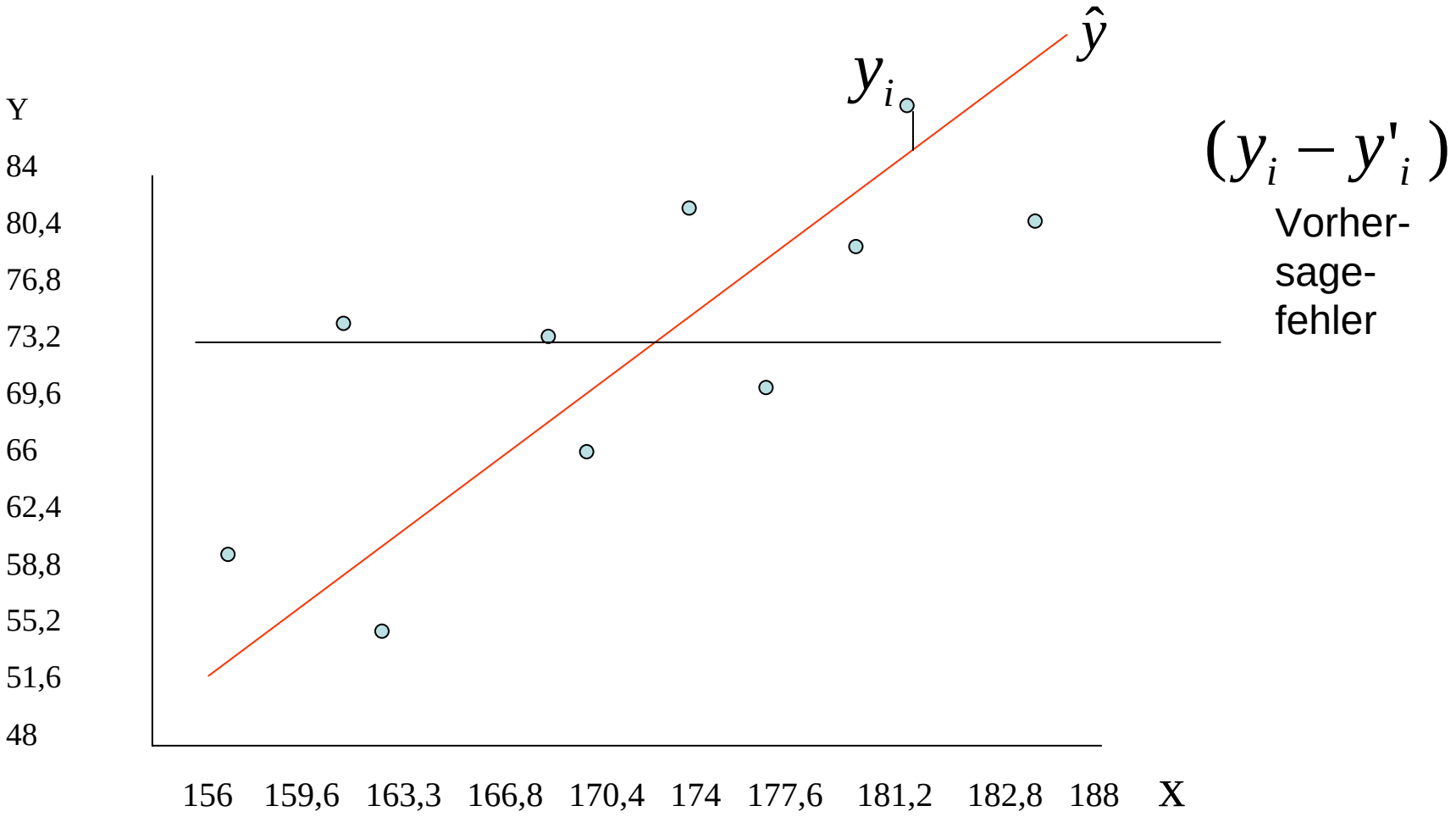
Streudiagramm mit den Abweichungen von der Regressionslinie = Nichterklärte Abweichungen



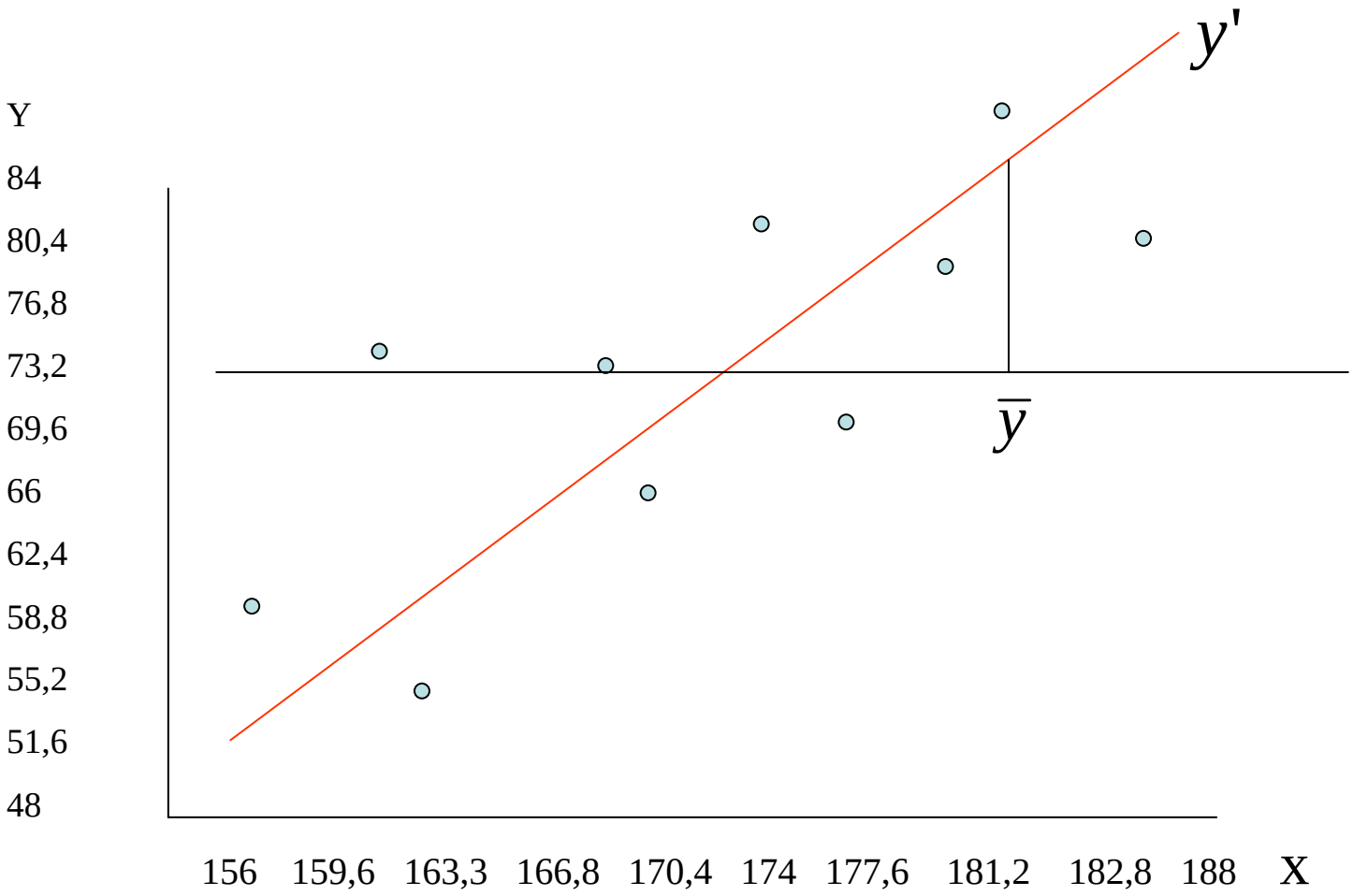
Streudiagramm mit der Abweichung vom Mittelwert



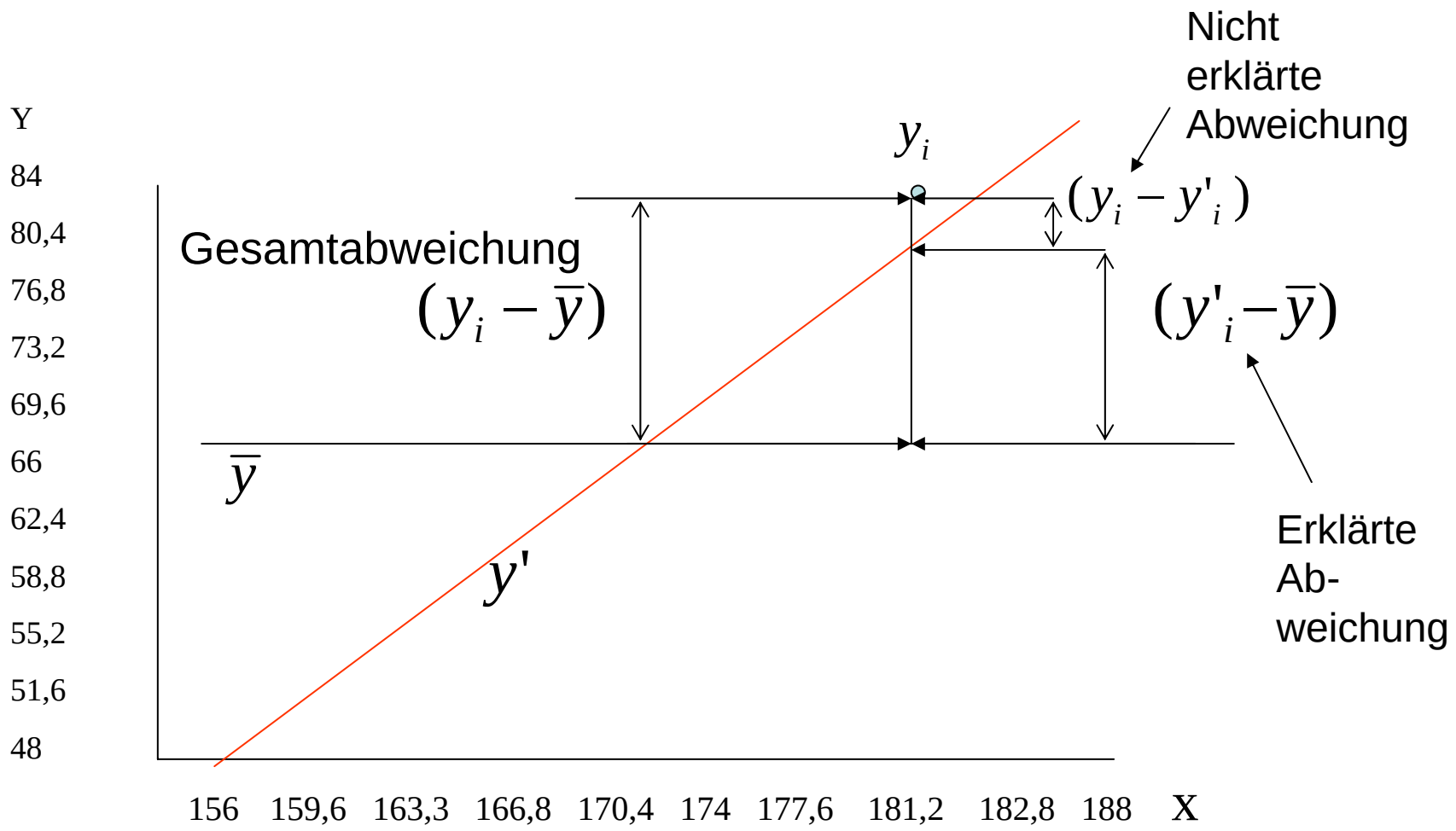
Streudiagramm mit der nicht erklärten Abweichung



Streudiagramm mit der erklärten Abweichung $(y'_i - \bar{y})$



Gesamtabweichung = Erklärte Abweichung + Nicht erklärte Abweichung



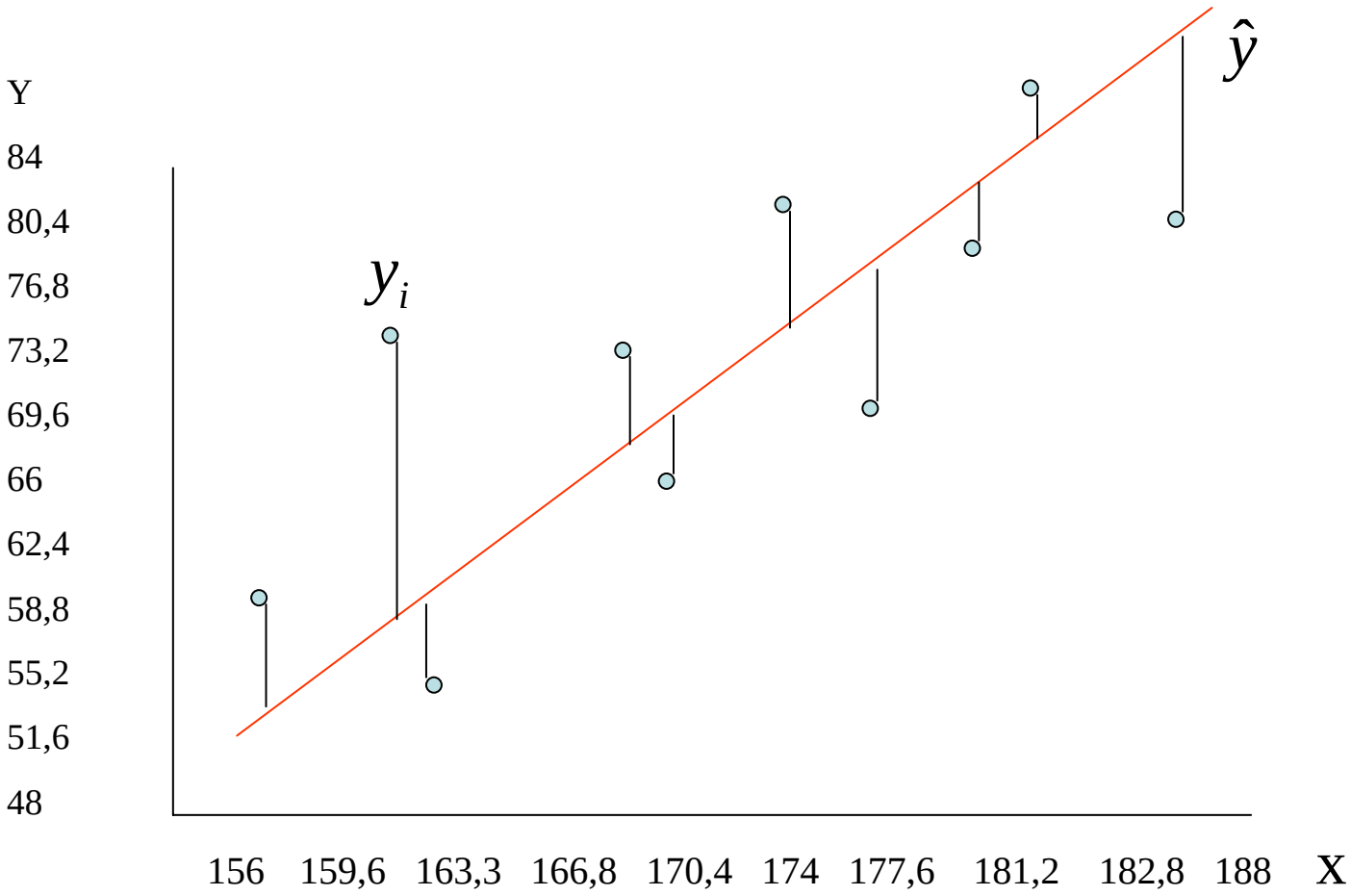
$$(y_i - \bar{y}) \quad = (y_i - y'_i) \quad + \quad (y'_i - \bar{y})$$

Die Regressionsgerade

Die Regressionsgerade ist die Gerade, welche den Punkteschwarm im Streudiagramm am besten repräsentiert.

Die Lokalisierung der Gerade erfolgt über die Methode der kleinsten Quadrate. Danach ist die Gerade so lokalisiert, dass die Summe der vertikalen Abweichungen der empirischen Werte von der Geraden gleich Null und die Summe der quadrierten Abweichungen ein Minimum ist.

Streudiagramm mit den Abweichungen von der Regressionslinie = Nichterklärte Abweichungen



Die Gleichung der Regressionsgeraden

$$y' = a + bx$$

a ist der Schnittpunkt mit der Ordinate
b die Steigung der Geraden

$$b_{yx} = \frac{\operatorname{cov}(x, y)}{s_x^2} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$$a_{yx} = \bar{y} - b_{yx} x$$

Berechnung von a_{yx} und b_{yx}

Körpergröße in cm x_i	Körpergewicht in kg y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
160	55	-15,2	231,04	-14,4	218,88
163	57	-12,2	148,84	-12,4	151,28
165	70	-10,2	104,04	0,6	-6,12
169	67	-6,2	38,44	-2,4	14,88
176	48	0,8	0,64	-21,4	-17,12
176	72	0,8	0,64	2,6	3,4
178	63	2,8	7,84	-6,4	-17,92
183	87	7,8	60,84	17,6	137,28
190	90	14,8	210,04	20,6	304,88
192	85	16,8	282,24	15,6	262,08
$\Sigma = \mathbf{1752}$	$\Sigma = \mathbf{694}$	$\Sigma = \mathbf{0}$	$\Sigma = \mathbf{1084,6}$	$\Sigma = \mathbf{0}$	$\Sigma = \mathbf{1051,52}$

$$b_{yx} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1752}{10} = 175,2 \qquad \bar{y} = \frac{694}{10} = 69,4$$

$$b_{yx} = \frac{1051,52}{1084,6} = 0,97$$

$$a_{yx} = 69,4 - 0,97(175.2) = -100,54$$

Der Determinationskoeffizient r^2

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{\text{Gesamtvariation - Nicht erklärte Variation}}{\text{Gesamtvariation}}$$

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{\Sigma (y_i - \bar{y})^2 - \Sigma (y_i - y'_i)^2}{\Sigma (y_i - \bar{y})^2}$$

$$r^2 = \frac{\text{Erklärte Variation}}{\text{Gesamtvariation}} = \frac{\Sigma (y'_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2}$$

r^2 repräsentiert jenen Teil der Gesamtvariation der abhängigen Variablen, der durch die unabhängige Variable erklärt wird

Berechnung von r^2

Körpergröße in cm x_i	Körpergewicht in kg y_i	y'_i	$y'_i - \bar{y}$	$(y'_i - \bar{y})^2$	$(y_i - \bar{y})^2$
160	55	54,66	-14,74	217,27	207,36
163	57	57,57	-11,83	139,94	153,76
165	70	59,51	-9,89	97,81	0,36
169	67	63,39	-6,01	36,12	5,76
176	48	70,18	0,78	0,60	457,96
176	72	70,18	0,78	0,60	6,76
178	63	72,12	2,72	7,39	40,96
183	87	76,97	7,57	57,30	309,76
190	90	83,76	14,36	206,21	424,36
192	85	85,70	16,30	265,69	243,36
$\Sigma = \mathbf{1752}$	$\Sigma = \mathbf{694}$	$\Sigma \approx \mathbf{694}$	$\Sigma \approx \mathbf{0}$	$\Sigma = \mathbf{1028,93}$	$\Sigma = \mathbf{1850,4}$

$$y' = -100,54 + 0,97x$$

$$r^2 = \frac{\text{Erklärte Variation}}{\text{Gesamtvariation}} = \frac{\Sigma(y'_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$r^2 = \frac{1028,93}{1850,4} = 0,556$$

$$1 - r^2 = 1 - 0,556 = 0,444$$

$1-r^2$ ist der Koeffizient der Nichtdetermination

$$1 = r^2 + 1-r^2$$

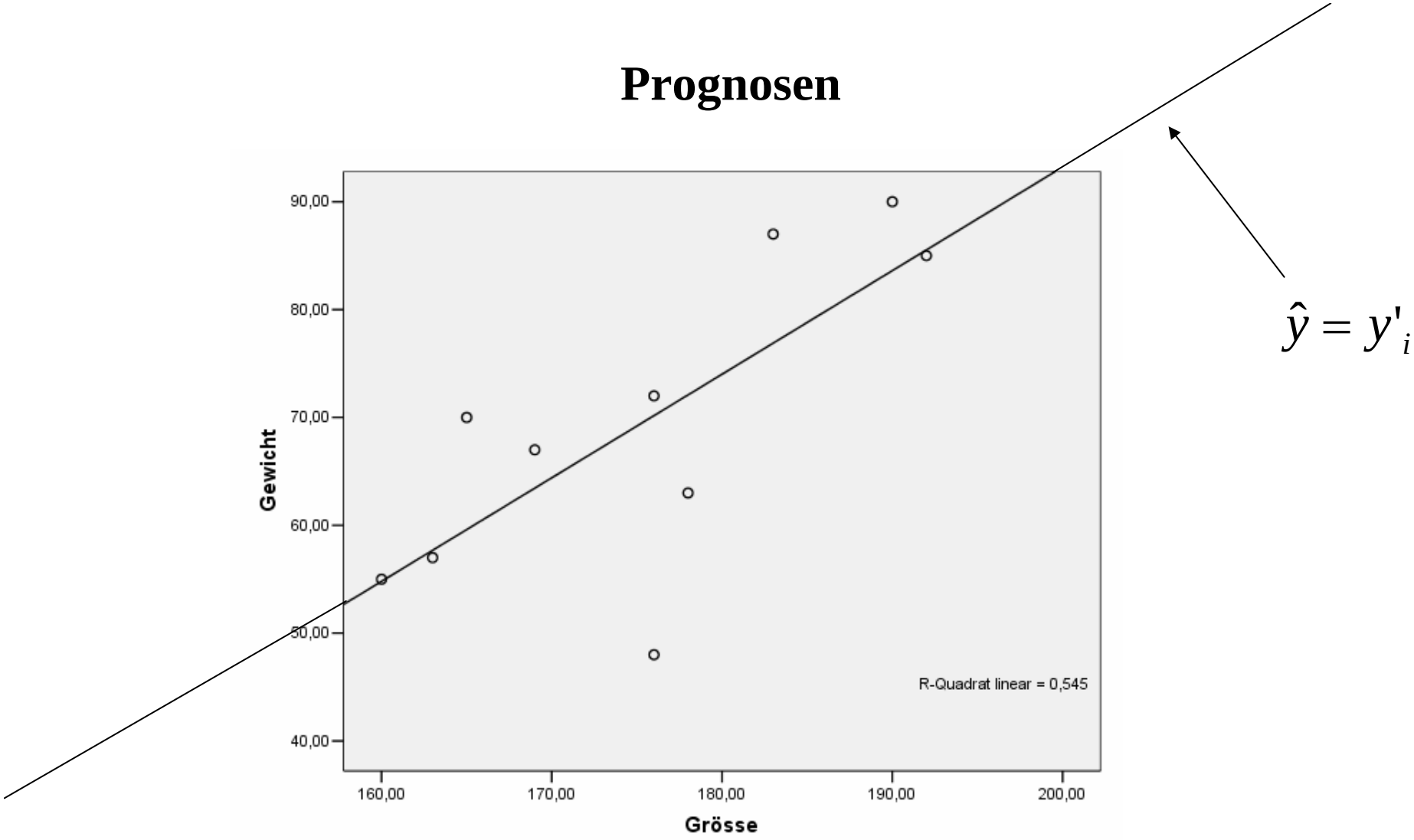
Gesamtvariation = Erklärte Variation + Nicht erklärte Variation

$$\Sigma(y_i-\bar{y})^2 = \Sigma(y'_i-\bar{y})^2 + \Sigma(y_i-y'_i)^2$$

Beziehung zwischen r, r^2 und $1-r^2$

Korrelations-koeffizient Pearsons r	**Determinations-koeffizient** r^2	**Koeffizient der Nichtdetermination** $1-r^2$
.10	.01	.99
.20	.04	.96
.30	.09	.91
.40	.16	.84
.50	.25	.75
.60	.36	.64
.70	.49	.51
.80	.64	.36
.90	.81	.19

Prognosen



$\hat{y} = y'_i$

$y' = -100{,}54 + 0{,}97x$

$y' = -100{,}54 + 0{,}97(180)$

$y' = 74{,}06$

Vorhersagen sind nur im Rahmen der Datengrundlage zulässig!