REGRESSIONSANALYSE

DAVID BUCHATZ
UNIVERSITÄT ZU KÖLN

AUFBAU

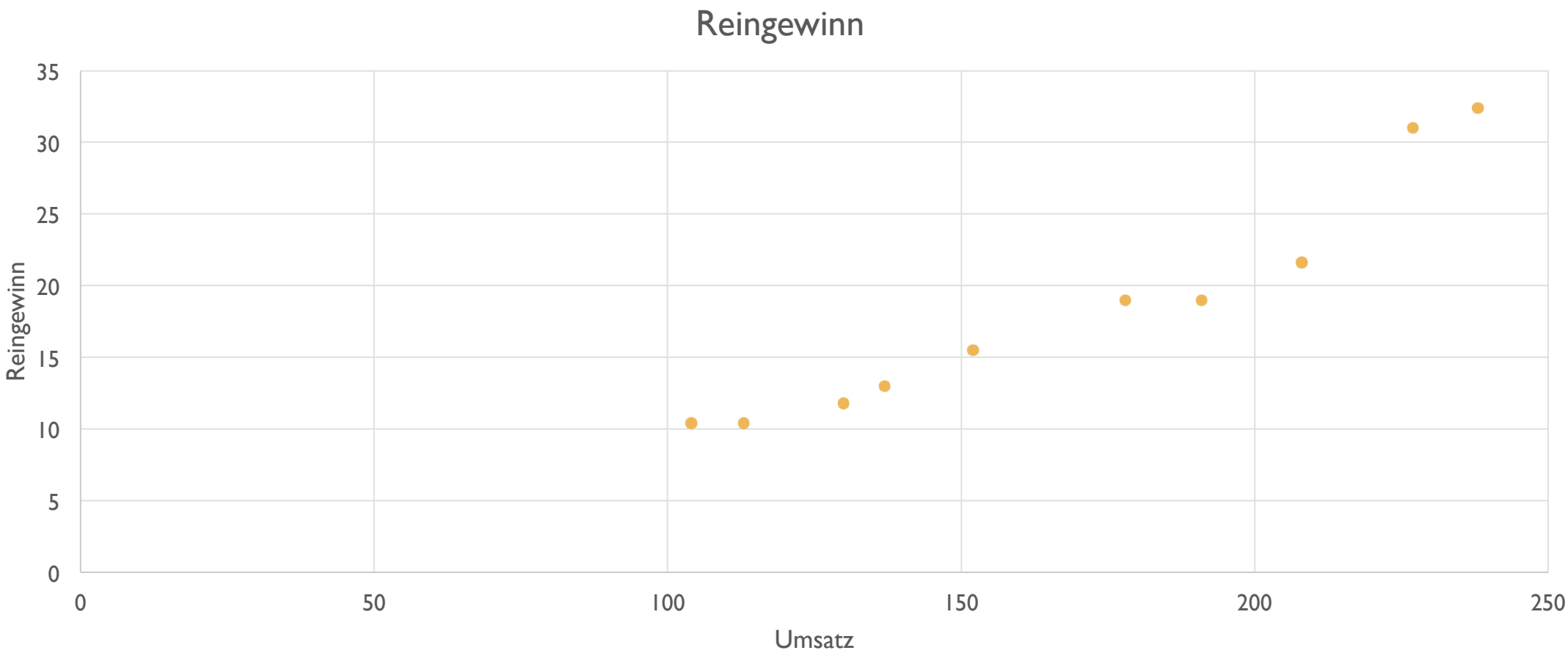
- §1 Historie
- §2 Anwendungen / Ziele
- §3 Lineare Regression/ Beispiel KQ
- §4 Nichtlineare Regression
- §5 Eigenschaften der Schätzer

Historie

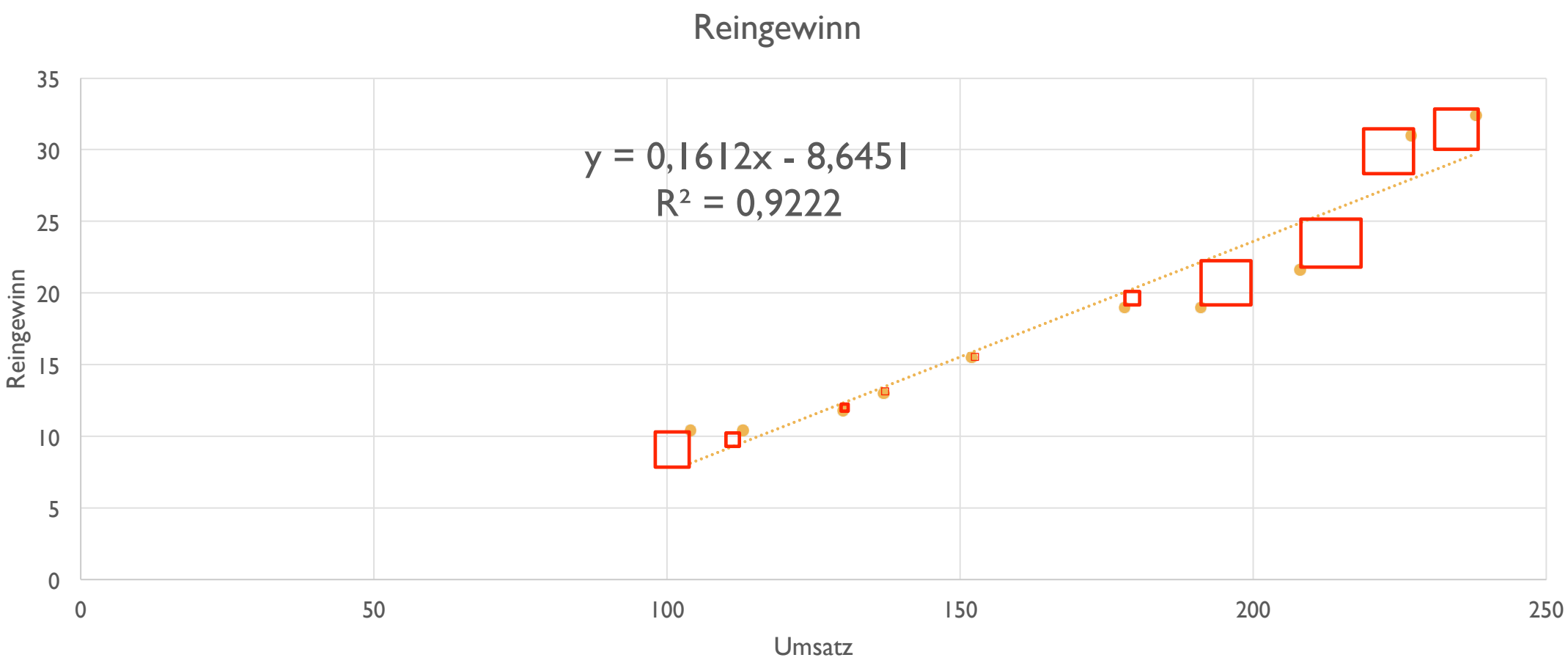
- früheste Form der Regression Methode der kleinsten Quadrate (frz.: *méthode des moindres carrés*), 1805 von Legendre und 1809 von Gauß veröffentlicht

- Umlaufbahnen der Planeten um die Sonne anhand von astronomischen Beobachtungen zu bestimmen

DATENPUNKTWOLKE



UNSER ZIEL


Reingewinn
y = 0,1612x - 8,6451
R² = 0,9222
Reingewinn
Umsatz
0
5
10
15
20
25
30
35
0
50
100
150
200
250

- 1 Unternehmen
- 10 verschiedene Filialen
- Einheitliches Konzept außer Aktionen & Rabatte
- Herr Uli Hoeneß stellt sich nun die Frage :

Wächst der Reingewinn bei Filialen mit größeren Umsätzen auch stetig an , oder verringert sich der Reingewinn bei höherem Umsatz durch irgendwelche Faktoren, die man dann untersuchen müsste
(z.B. kleinere Gewinnmarge an Produkten und eventuell höhere Kosten, bzw. falsches Management) ?

ANWENDUNGEN

- Regressionsverfahren haben viele **praktische Anwendungen**

- **Prognose oder Vorhersage**
- **Stärke des Zusammenhangs**

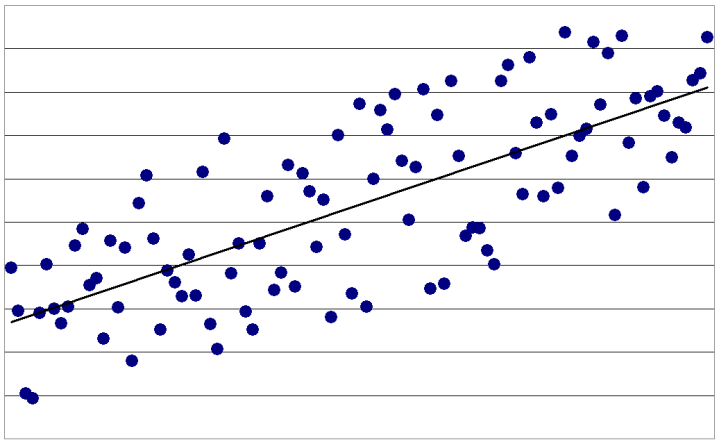
- Wenn nun zusätzliche Werte *x* ohne zugehörigen Wert *y* vorliegen, dann kann das angepasste Modell zur Vorhersage des Wertes von *y* verwendet werden.

SCHEMA EINER REGRESSIONSANALYSE

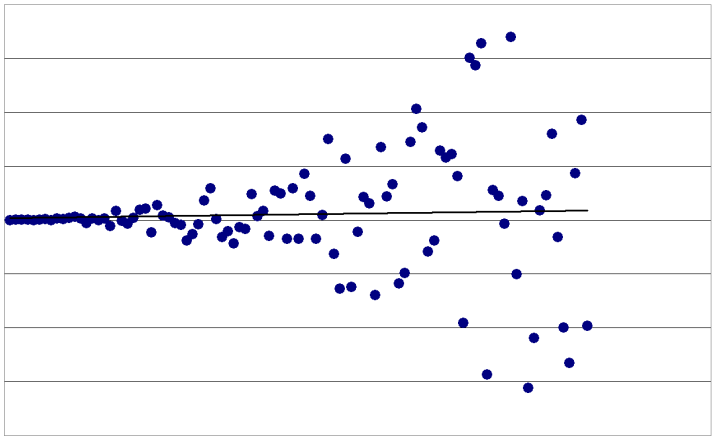
- Am **Beginn** jedes statistischen Verfahrens steht die **Aufbereitung der Daten**, insbesondere :
- **Plausibilisierung**
- der Umgang mit **fehlenden Daten**.

- **<u>linearen Regression</u>**
  linearer Zusammenhang sowie **Homoskedastizität** wird vorausgesetzt.

Homoskedastizität



Heteroskedastizität



- die Berücksichtigung von **Interaktionen** (bei linearer Regression)
- *Untersuchung der Daten auf* **Ausreisser** *und* **einflussreiche Datenpunkte**

MATHEMATISCHE FORMULIERUNG

- Beziehung zwischen den unabhängigen Variablen x und der abhängigen Variable y :

- y= f(x) + e im 1-dimensionalen Fall

- y= f($x\downarrow 1$, $x\downarrow 2$,…., $x\downarrow n$) + e im n-dimensionalen Fall

- f die gesuchte oder angenommene Funktion
- e den Fehler bzw. das Residuum des Modells

LINEARE REGRESSION

- lineare Regression Modell so spezifiziert, dass abhängige Variable y als Linearkombination der Parameter (= Regressionskoeffizient) $\beta \downarrow i$ dargestellt wird , und nicht von der unabhängigen Variablen x.
- Zur Bestimmung der Modellparameter $\beta \downarrow i$ wird die Methode der kleinsten Quadrate eingesetzt.

HERLEITUNG UND VERFAHREN

- Eine einfache Modellfunktion mit zwei linearen Parametern stellt das Polynom erster Ordnung

$$f(x) = \beta_0 + \beta_1 * x$$

- zu n gegeben Messwerten $(x_1 , y_1),\ldots..(x_n , y_n)$ die Koeffizienten β_0 und β_1 der bestangepassten Gerade

- Die **Abweichungen $r\downarrow i$** zwischen der gesuchten Geraden und den jeweiligen Messwerten

$$
\begin{aligned}
r\downarrow 1 &= \beta\downarrow 0 + \beta\downarrow 1\, x\downarrow 1 - y\downarrow 1 \\
r\downarrow 2 &= \beta\downarrow 0 + \beta\downarrow 1\, x\downarrow 2 - y\downarrow 2 \\
&\ldots \\
r\downarrow n &= \beta\downarrow 0 + \beta\downarrow 1\, x\downarrow n - y\downarrow n
\end{aligned}
$$

nennt man **Anpassungsfehler** oder **Residuen**.
Gesucht sind nun die Koeffizienten $\beta\downarrow 0$ und $\beta\downarrow 1$ mit der **kleinsten Summe der Fehlerquadrate.**

- $\min_{\beta_0, \beta_1} \sum_{i=1}^{n} r_i^2$

- Vorteil des Ansatzes Quadrat der Fehler wird sichtbar, wenn man Minimierung mathematisch durchführt : Die Summenfunktion wird als Funktion der beiden Variablen β_0 und β_1 aufgefasst
- Dann partielle Ableitung mit folgendem LGS

$$n*\beta_0 + \left(\sum_{i=1}^{n} x_i\right) * \beta_1 = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right) * \beta_0 + \left(\sum_{i=1}^{n} x_i^2\right) * \beta_1 = \sum_{i=1}^{n} x_i y_i$$

- Nullstelle suchen (Minumum)

- Mit der Lösung

$$\beta\downarrow 1 = (\sum i=1\uparrow n \blacksquare x\downarrow i\uparrow \, y\downarrow i\uparrow) - n*(x\,y)\,/(\sum i=1\uparrow n \blacksquare x\downarrow i\uparrow 2\ ) - n*(x)^2$$

und

$$\beta\downarrow 0 = y - \beta\downarrow 1\, x$$

- Dabei ist

$$x = 1/n * \sum i=1\uparrow n \blacksquare x\downarrow i\uparrow$$

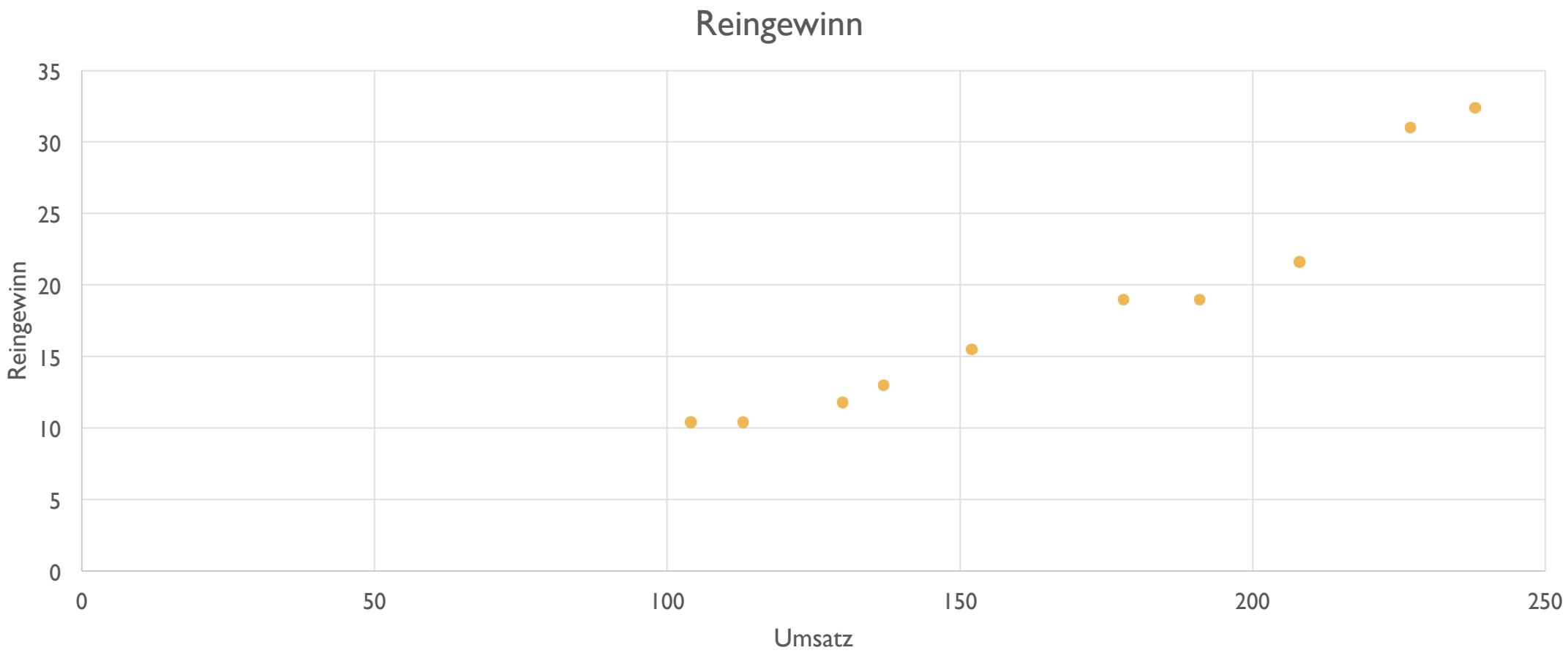
das arithmetische Mittel der x-Werte, y entsprechend

- Die Lösung für $\beta\downarrow 1$ kann mit Hilfe des Verschiebungssatzes auch als

$$\beta\downarrow 1 = \sum i=1\uparrow n\blacksquare(x\downarrow i\uparrow - x)(y\downarrow i\uparrow - y)/\sum i=1\uparrow n\blacksquare(x\downarrow i\uparrow - x)^2$$

angegeben werden.

BERECHNUNG DER PARAMETER



Nummer	Umsatz	Reingewinn	$x_i - x$	$y_i - y$			
i	x_i	y_i	$x_i^* - x^*$	$y_i^* - y^*$	$x_i^* - x^* * y_i^* - y^*$	$x_i^* - x^* * x_i^* - x^*$	$y_i^* - y^* * y_i^* - y^*$
1	208	21,6	40,2	3,19	128,238	1616,04	10,1761
2	152	15,5	−15,8	−2,91	45,978	249,64	8,4681
3	113	10,4	−54,8	−8,01	438,948	3003,04	64,1601
4	227	31,0	59,2	12,59	745,328	3504,64	158,5081
5	137	13,0	−30,8	−5,41	166,628	948,64	29,2681
6	238	32,4	70,2	13,99	982,098	4928,04	195,7201
7	178	19,0	10,2	0,59	6,018	104,04	0,3481
8	104	10,4	−63,8	−8,01	511,038	4070,44	64,1601
9	191	19,0	23,2	0,59	13,688	538,24	0,3481
10	130	11,8	−37,8	−6,61	249,858	1428,84	43,6921
Σ	1678	184,1	0,0	0,00	3287,820	20391,60	574,8490

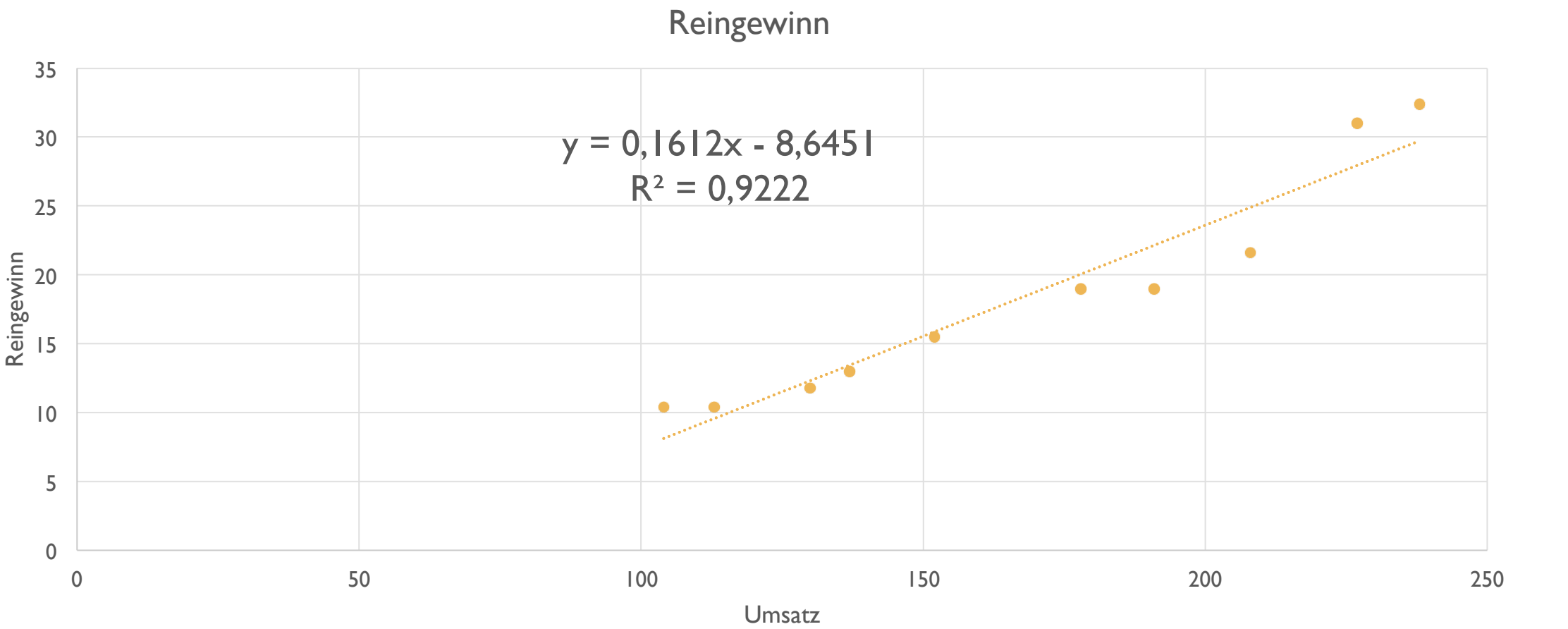
- Dann bestimmt man $\beta\downarrow 1$:

$$\beta\downarrow 1 = \sum i=1\uparrow n\blacksquare (x\downarrow i\uparrow - x)(y\downarrow i\uparrow - y)/\sum i=1\uparrow n\blacksquare (x\downarrow i\uparrow - x)^2 = 3287{,}820/20391{,}60 = \mathbf{0{,}1612}$$

mit jeder Millionen € Umsatz wächst Reingewinn im Durchschnitt etwa
0,16 Mio. €

- Das Absolutglied $\beta\downarrow 0$ ergibt sich als :

- $\beta\downarrow 0 = y - \beta\downarrow 1\, x$ = 18,41 – 0,1612 * 167,8 = -8,6451


Reingewinn
y = 0,1612x - 8,6451
R² = 0,9222
Reingewinn
Umsatz
0
5
10
15
20
25
30
35
0
50
100
150
200
250

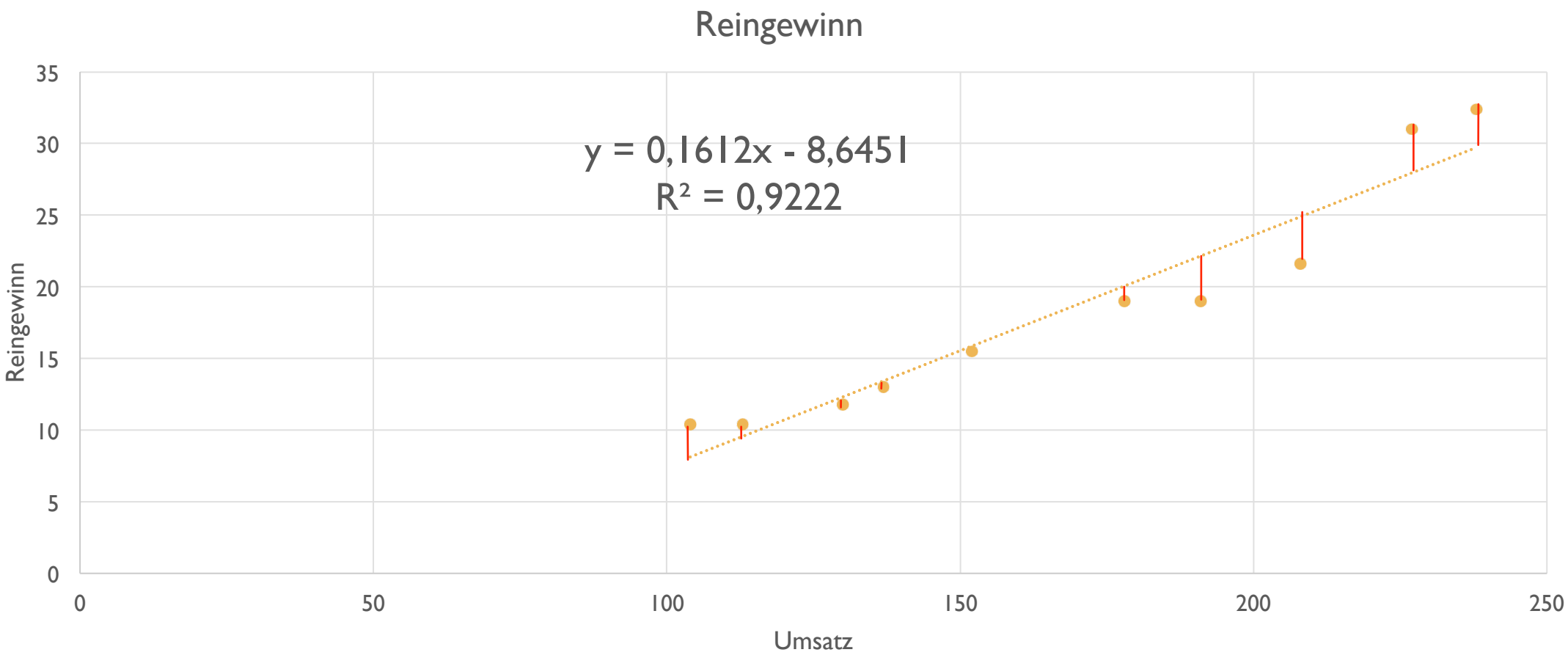
- Anpassung der Punkte recht gut
- Im Mittel beträgt die Abweichung 2,1 Mio. €
- Bestimmtheitsmaß ergibt einen Wert von ca. 92%

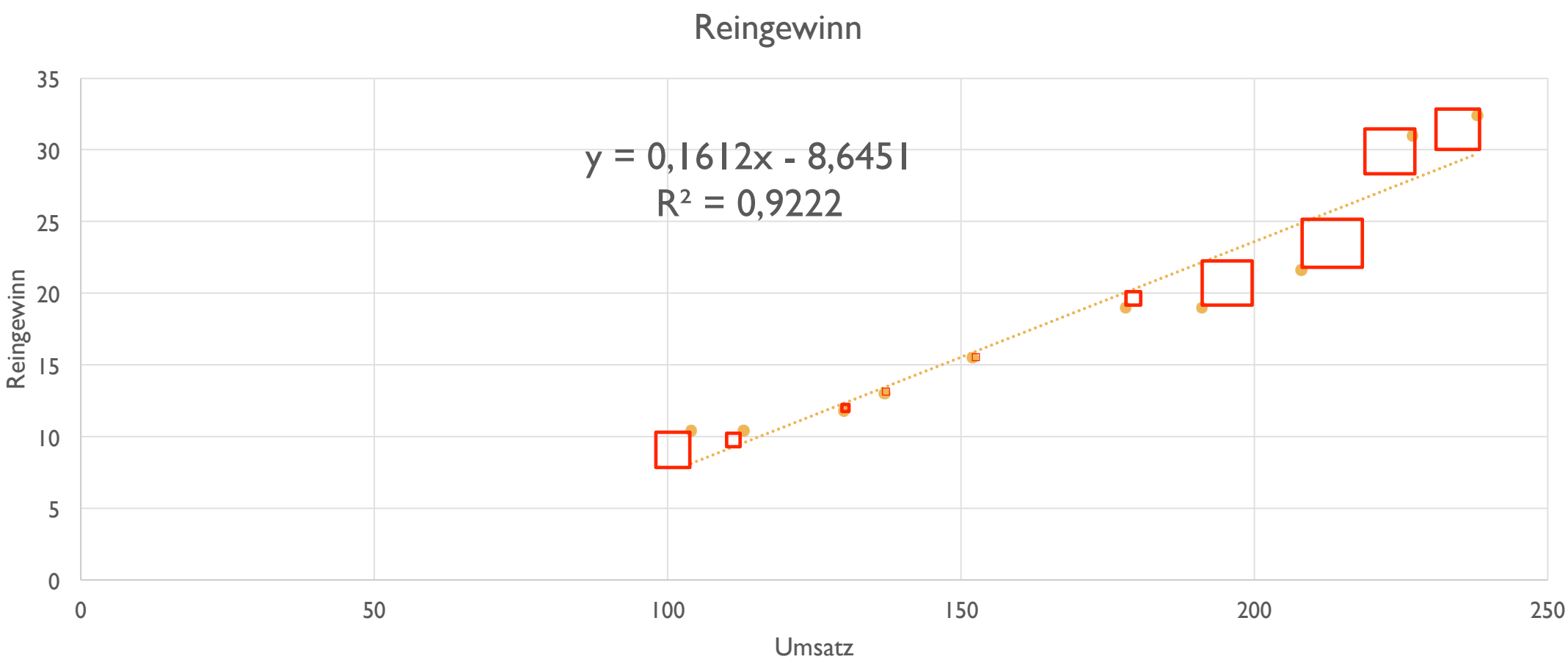
- **Bestimmtheitsmaß** (auch **Determinationskoeffizient**) Maß in der Statistik für den erklärten Anteil der Variabilität (Varianz) einer abhängigen Variablen durch ein statistisches Modell.
- Indirekt Zusammenhang zwischen der abhängigen und den unabhängigen Variablen

- 0 (oder 0 %): kein linearer Zusammenhang und
- 1 (oder 100 %): perfekter linearer Zusammenhang.

- $R^2 = 1 - \sum_{i=1}^{n} r_i^2 \; / \sum_{i=1}^{n} (y_i - \bar{y})$

RESIDUEN


Reingewinn
y = 0,1612x - 8,6451
R² = 0,9222
Reingewinn
Umsatz
0
5
10
15
20
25
30
35
0
50
100
150
200
250

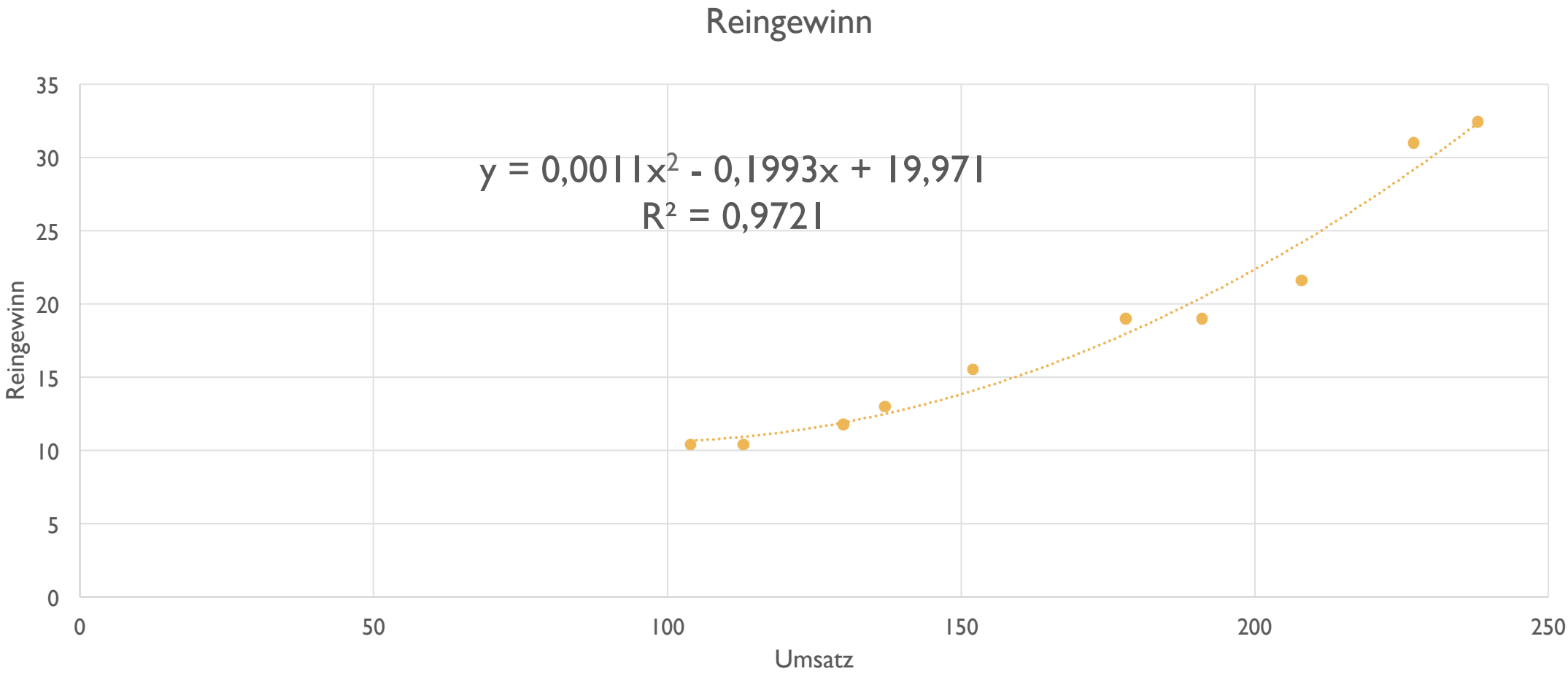

Reingewinn
y = 0,1612x - 8,6451
R² = 0,9222
Reingewinn
Umsatz
0
5
10
15
20
25
30
35
0
50
100
150
200
250

WARUM SUMME DER FEHLERQUADRATE ?

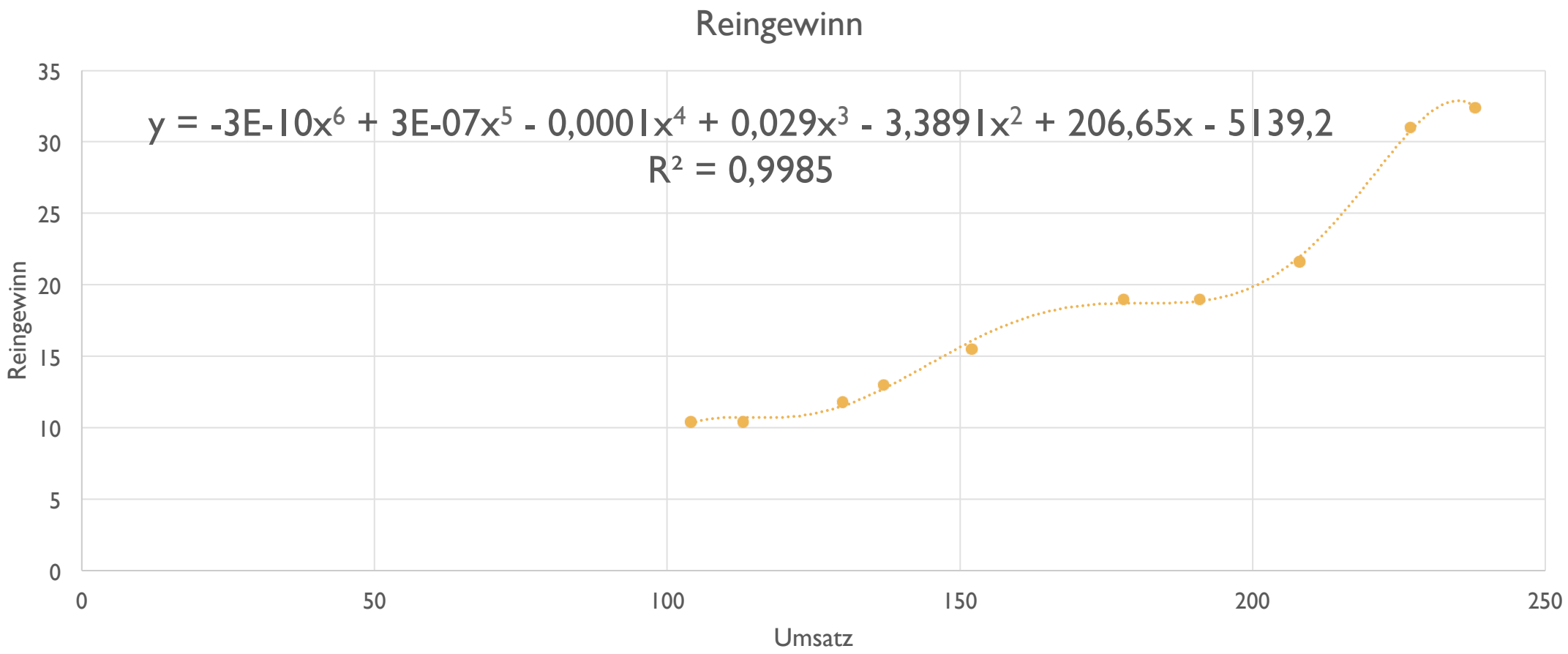
- das Vorzeichenproblem der Residuen hat man damit gelöst
- Die Summe der Residuen – Quadrate ist minimal (per Extremwertproblem gelöst)
- Bei jeder anderen Gerade wären vielleicht einzelne Residuen-Quadrate kleiner
- Jedoch Summe dieser Residuen-Quadrate größer als bei der eben ermittelten Geraden

- Allgemeiner als eine lineare Ausgleichsgerade sind Ausgleichspolynome
- $y(x) \approx \beta\downarrow 0 + \beta\downarrow 1 * x + \beta\downarrow 2 * x^2 + \ldots + \beta\downarrow q\, x\uparrow q$
- Lineare Regression, da man die einzelnen $\beta\downarrow i$ bestimmen will

BEISPIEL POLYNOM 2. GRADES



BEISPIEL POLYNOM 6. GRADES



NICHTLINEARE MODELLFUNKTIONEN

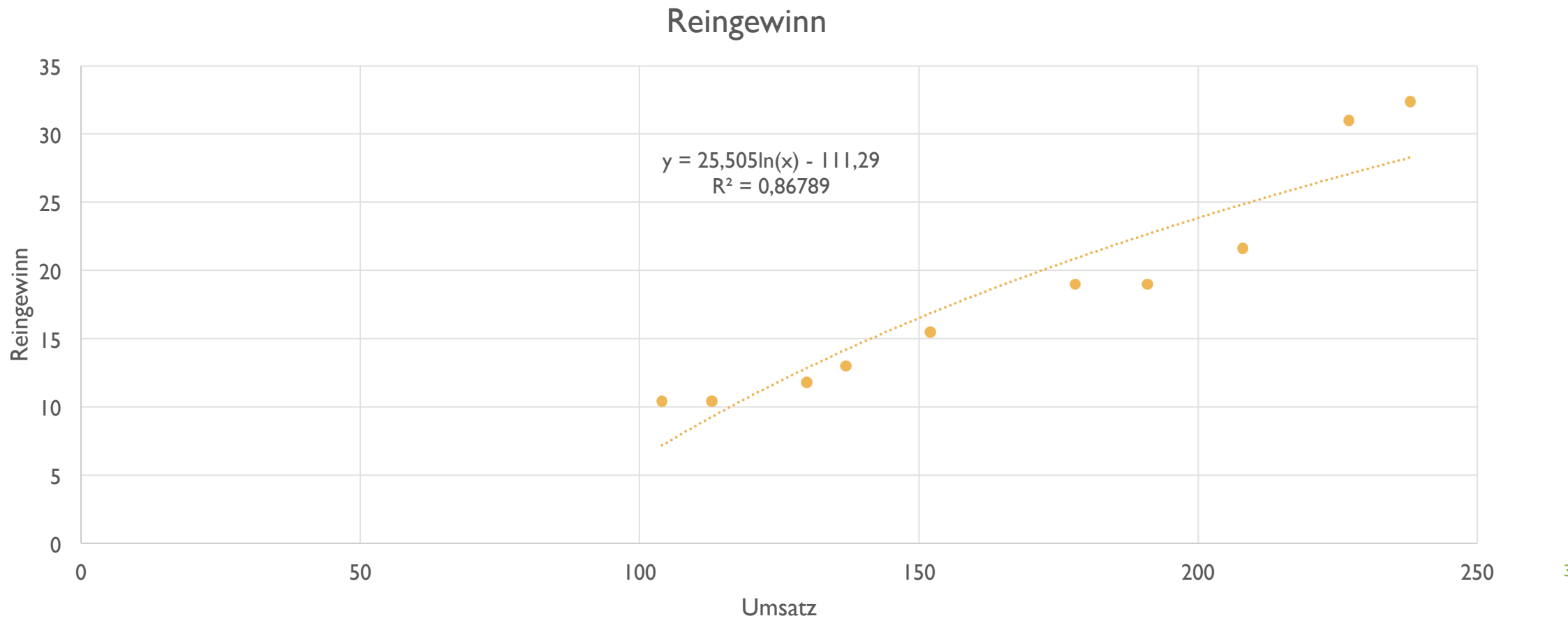
- leistungsfähige Rechner ermöglichen nichtlineare Regression
- Nichtlineare Modellierung ermöglicht Anpassung von Daten an jede Gleichung

- $y = f(\beta_{\downarrow i})$

- „curve fitting“
- Manche nichtlineare Probleme durch geeignete Substitution in lineare überführen
- multiplikatives Modell in der Form wie

$$y = \beta_0 * x^{\beta_1}$$

- lässt sich beispielsweise durch **Logarithmieren** in ein additives System überführen.
- Dieser Ansatz findet unter Anderem in der **Wachstumstheorie** Anwendung.


Reingewinn
y = 25,505ln(x) - 111,29
R² = 0,86789
Reingewinn
Umsatz
0
5
10
15
20
25
30
35
0
50
100
150
200
250

EIGENSCHAFTEN DER SCHÄTZER

- Seien $X\downarrow 1$, $X\downarrow 2$, …, $X\downarrow n$ eine Zufallsstichprobe der Größe n
- $\theta \downarrow n$ ein Schätzer auf Basis dieser Stichprobe und θ der wahre Parameter

UNVERZERRHEIT / ERWARTUNGSTREUE

unverzerrt /erwartungstreu (engl. : unbiased), wenn gilt :

$$E(\theta \downarrow n) = \theta$$

Weicht $E(\theta \downarrow n)$ von θ ab, dann verzerrt

Verzerrung Bias($\theta \downarrow n$)

$$\text{Bias}(\theta \downarrow n) = E(\theta \downarrow n) - \theta$$

asymptotische erwartungstreue Schätzfunktion muss nur gelten :

$$\lim_{n \to \infty} E(\theta \downarrow n) = \theta$$

KONSISTENZ / ASYMPTOTISCHE ERWARTUNGSTREUE

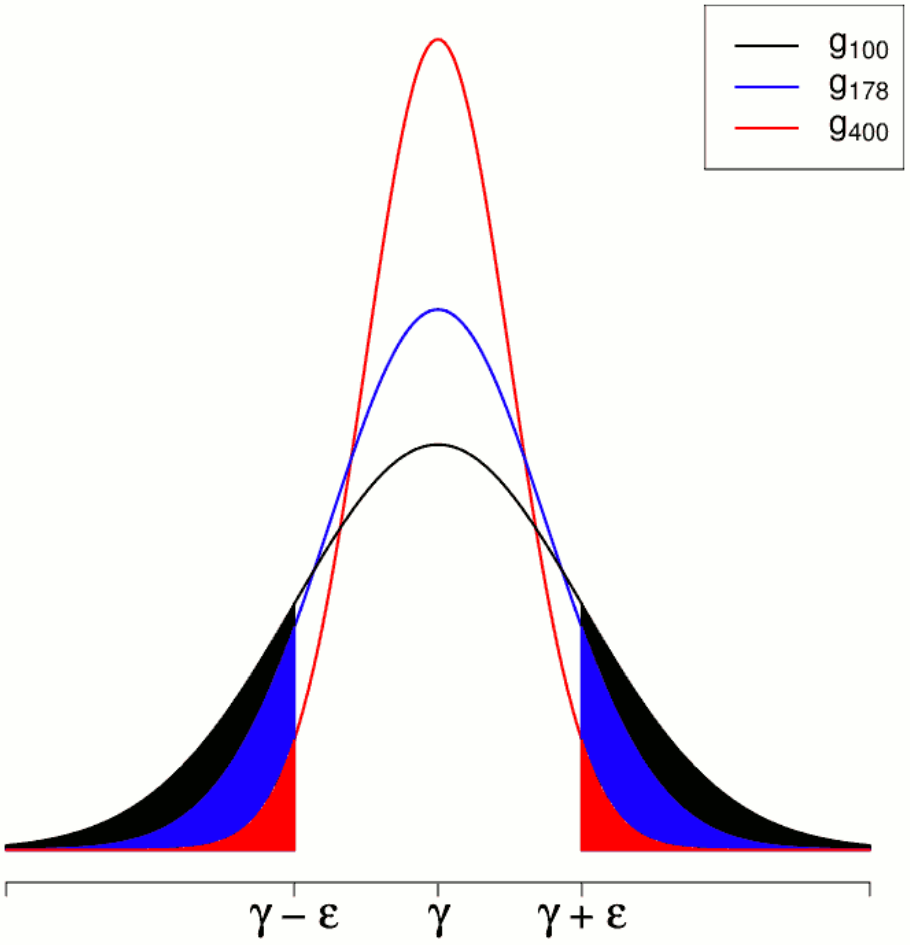
- Schätzer $\theta \downarrow n$ von θ ist konsistent / asymptotisch erwartungstreu (engl. : consistent), wenn für $\delta > 0$ gilt :

$$P\left(|\theta \downarrow n - \theta| > \delta \right) \rightarrow 0 \quad \text{für } n \rightarrow \infty$$

- stochastische Konvergenz
- Mit einfachen Worten : Eine konsistente Schätzfunktion nähert sich mit wachsendem n immer mehr dem wahren Parameter θ an (schätzt den wahren Parameter immer genauer)
- Konsistente Schätzfunktionen müssen daher **mindestens** asymptotisch erwartungstreu sein

- Wahrscheinlichkeitsdichten für die konsistenten Schätzfunktionen $\theta \downarrow n$ (n=100,178,400).

- Mit steigendem Stichprobenumfang wird der unbekannte Parameter θ immer genauer geschätzt.



MINIMALE VARIANZ , EFFIZIENZ

- Die Schätzfunktion soll eine möglichst kleine Varianz haben
- effiziente, beste oder wirksamste Schätzfunktion

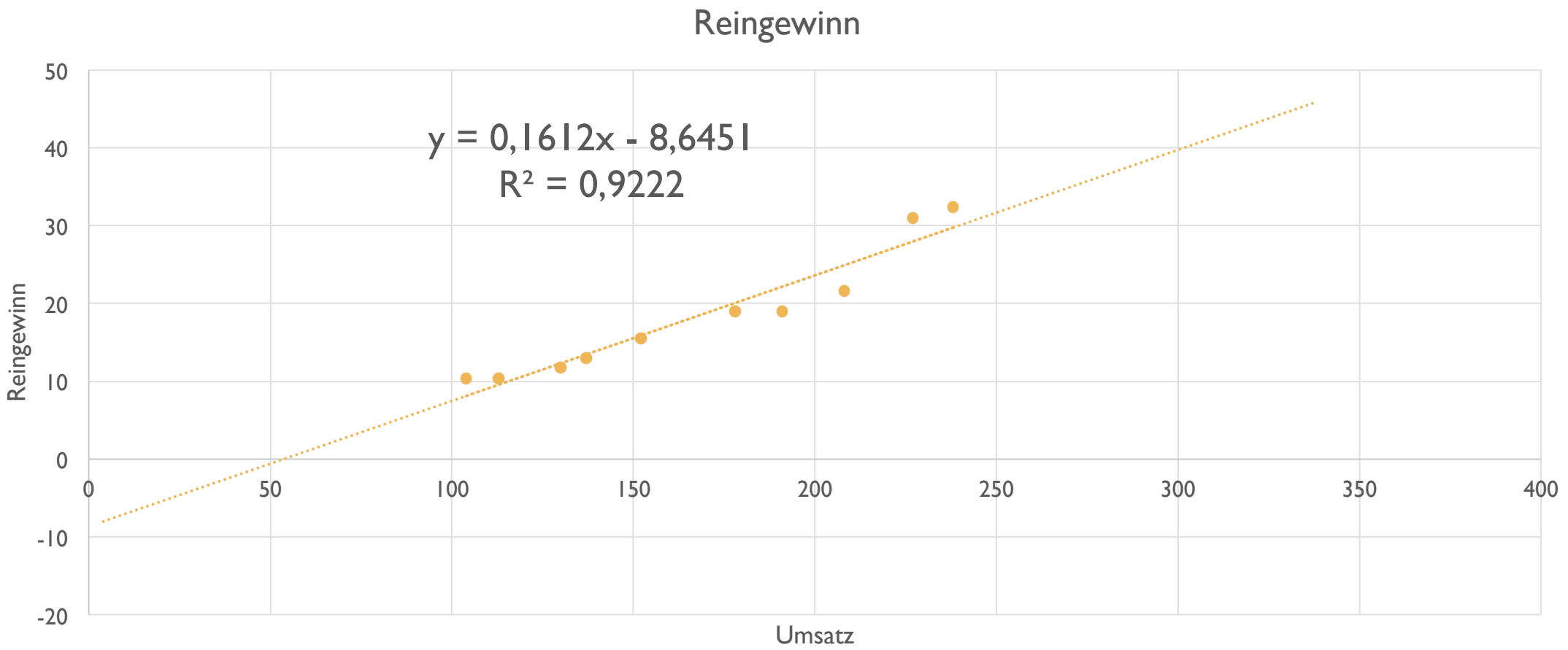
$$\mathrm{Var}(\theta_n^*) \leq \min_{\theta_n} Var(\theta_n)$$

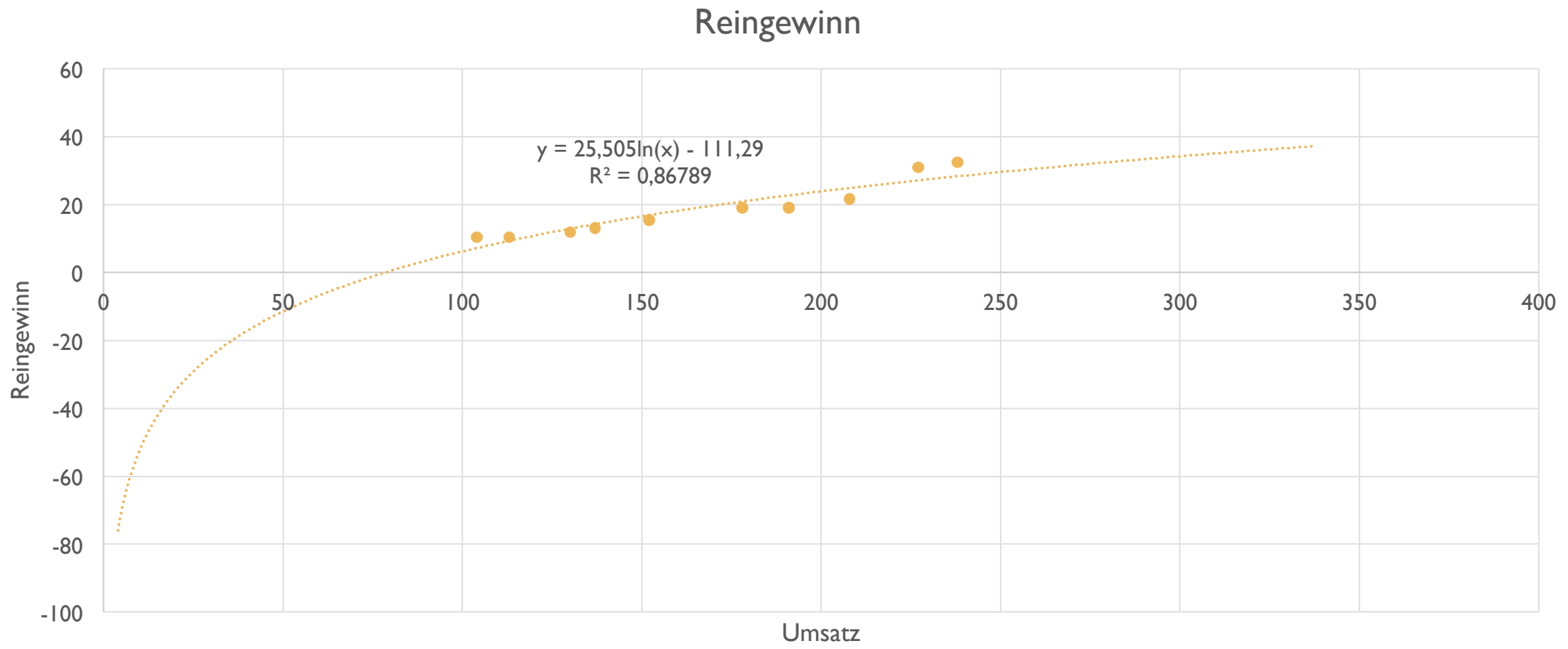
MITTLERER QUADRATISCHER FEHLER

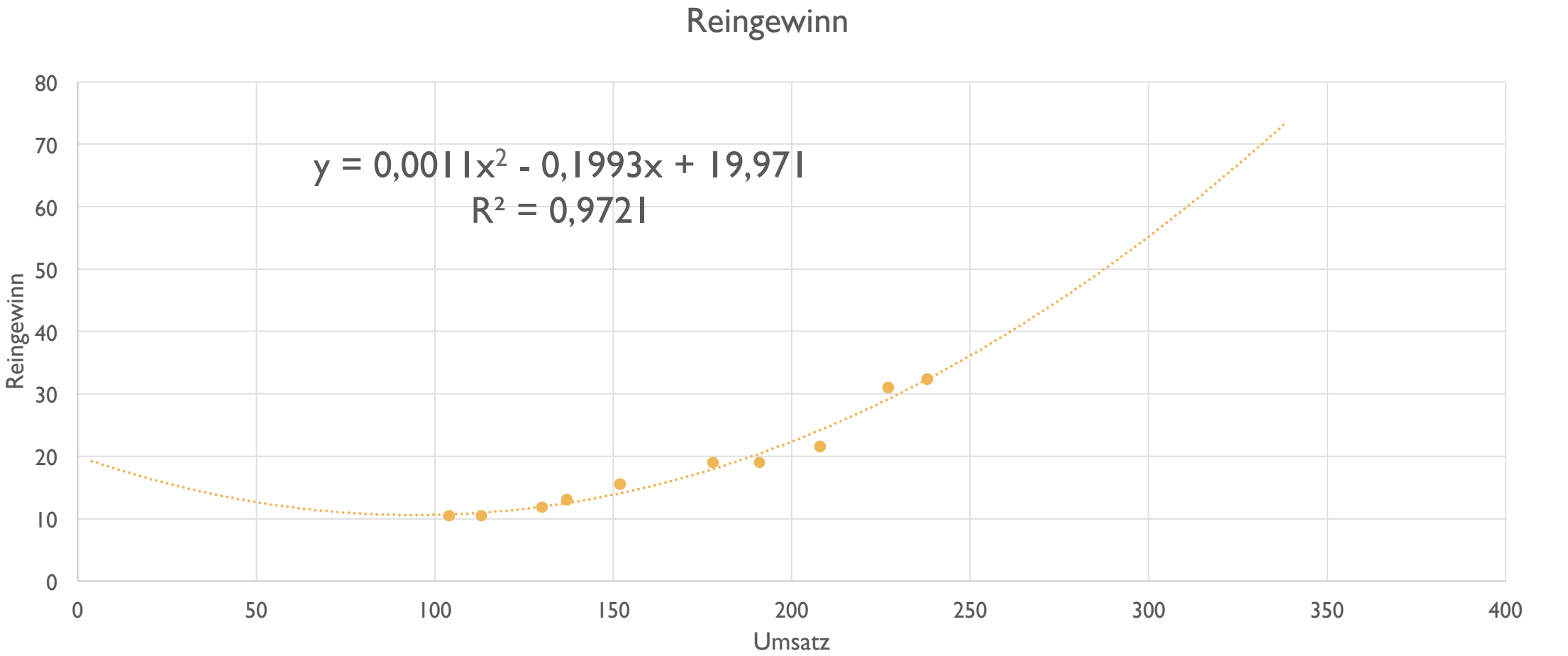
- Die Genauigkeit einer Schätzfunktion bzw. eines Schätzers wird oft durch seinen mittleren quadratischen Fehler (engl.: *mean squared error*) ausgedrückt
- möglichst kleinen mittleren quadratischen Fehler

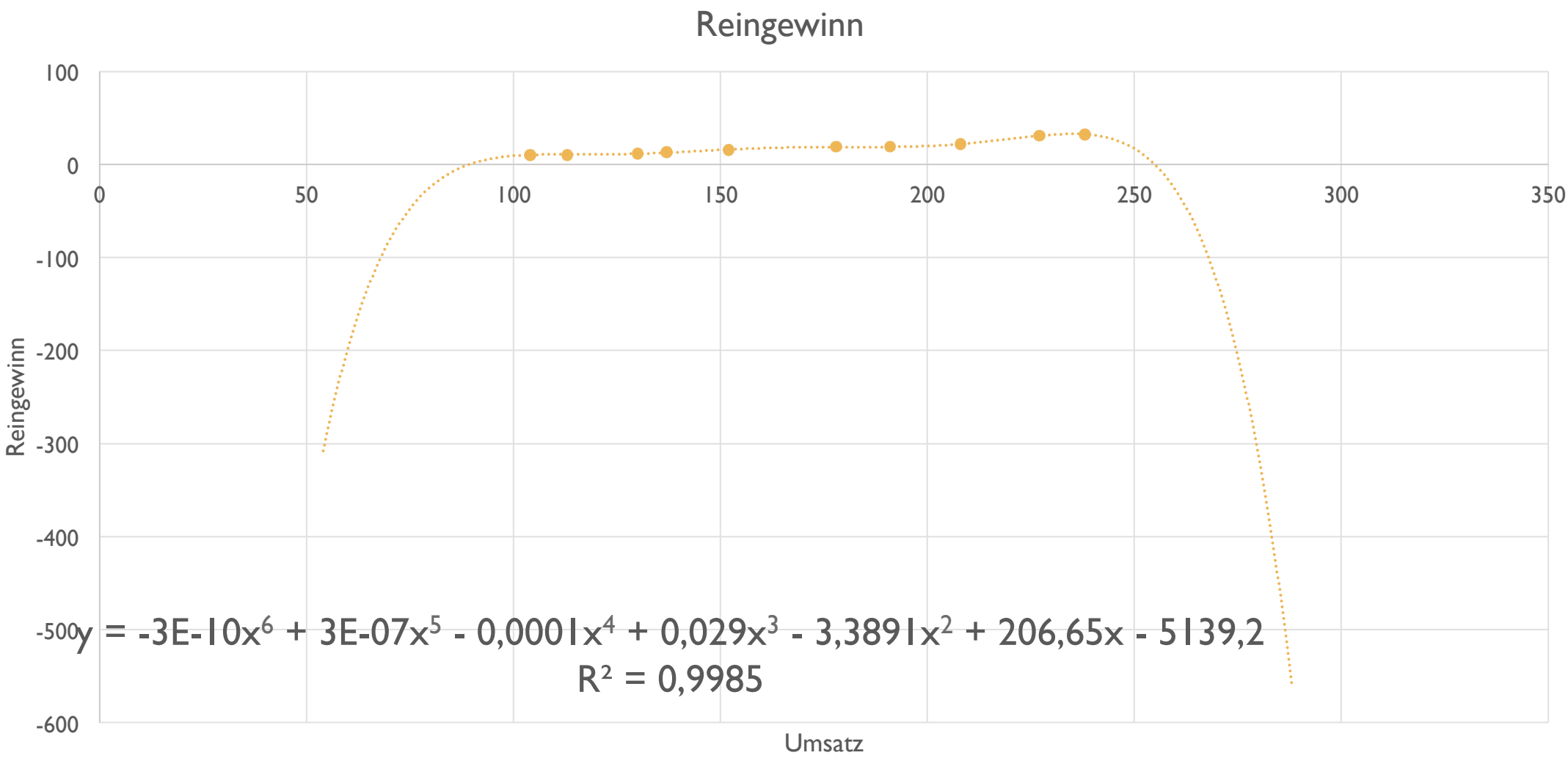
$$MSE(\theta_{\downarrow n}) = E\left[(\theta_{\downarrow n} - \theta)^2\right] = (Bias(\theta_{\downarrow n}))^2 + Var(\theta_{\downarrow n})$$

Prognosen→Sinn?




Reingewinn
y = 25,505ln(x) - 111,29
R² = 0,86789
60
40
20
0
-20
-40
-60
-80
-100
0
50
100
150
200
250
300
350
400
Reingewinn
Umsatz


Reingewinn
y = 0,0011x² - 0,1993x + 19,971
R² = 0,9721
Reingewinn
Umsatz
0
10
20
30
40
50
60
70
80
0
50
100
150
200
250
300
350
400


Reingewinn
100
0
-100
-200
-300
-400
-500
-600
0
50
100
150
200
250
300
350
Reingewinn
Umsatz
y = -3E-10x⁶ + 3E-07x⁵ - 0,0001x⁴ + 0,029x³ - 3,3891x² + 206,65x - 5139,2
R² = 0,9985