lineare Regression

Überblick

1. Einsatzgebiet
2. Vorhersage über Methode der kleinsten Quadrate
3. Regressionsgleichung
4. Voraussetzungen
5. Güte der Vorhersage
   - Standardschätzfehler
   - Konfidenzintervall
6. Kreuzvalidierung
7. Regression zur Mitte

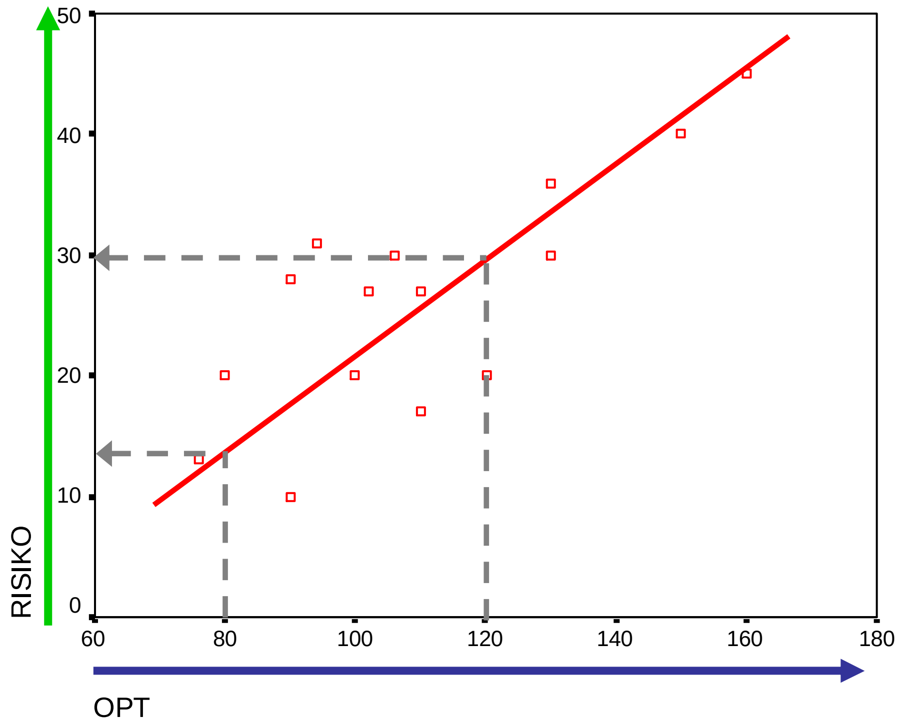
1. Sinn & Zweck

- Vorhersage einer Variable durch eine andere
  - vorherzusagende Variable (y) =
  - zur Vorhersage verwendete Variable (x) =
- Bsp: Vorhersage…
  - durch
  - durch

2. Vorhersage-Prinzip

- Ermittlung einer

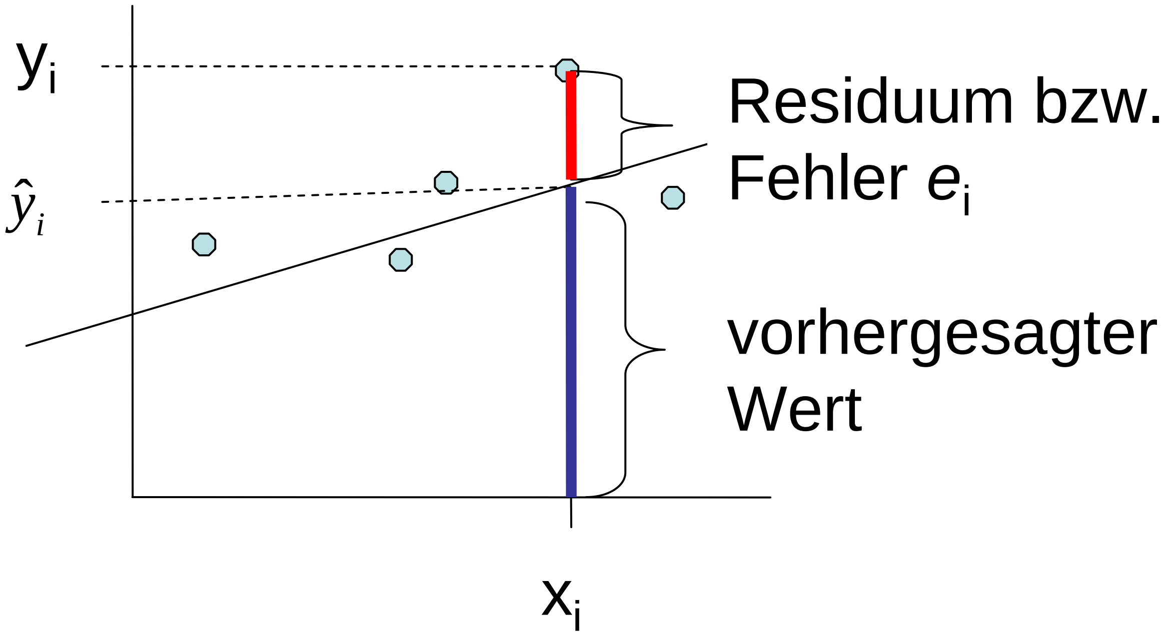
- nach der



Methode der kleinsten Quadrate

- Regressionsgerade wird so durch die Punktewolke gelegt, daß der quadrierte Vorhersagefehler über alle Probanden minimal ist:

$$\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 = min$$



3. Regressionsgleichung I

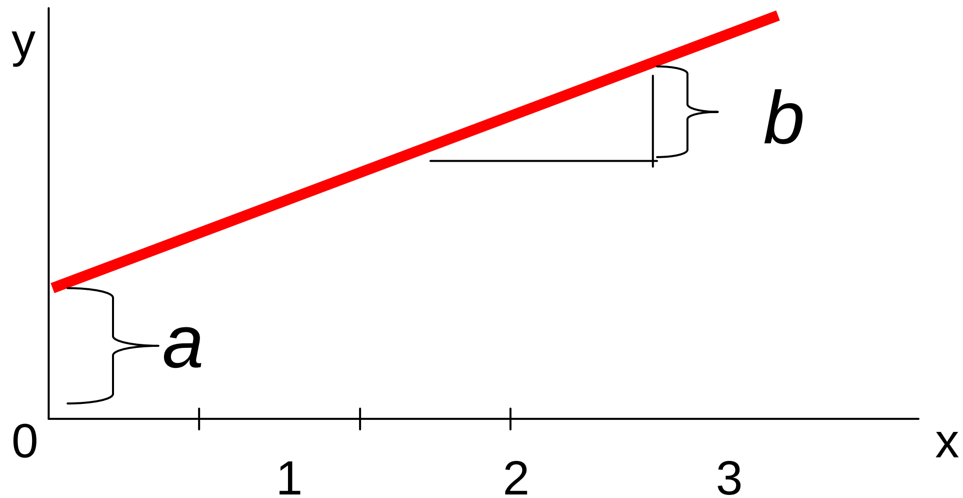
Regressionskoeffizient

Prädiktorwert

vorhergesagter Wert

$$\boxed{\hat{y}_i = b_{y.x} \cdot x_i + a_{y.x}}$$

Konstante



Beispiel

$$\boxed{\hat{y}_i} = 0.20 \cdot \boxed{x_i} + 1.00$$

Für Patientin mit einem Depressionswert von 20 soll der Ängstlichkeitswert geschätzt werden:

Regressionsgleichung II

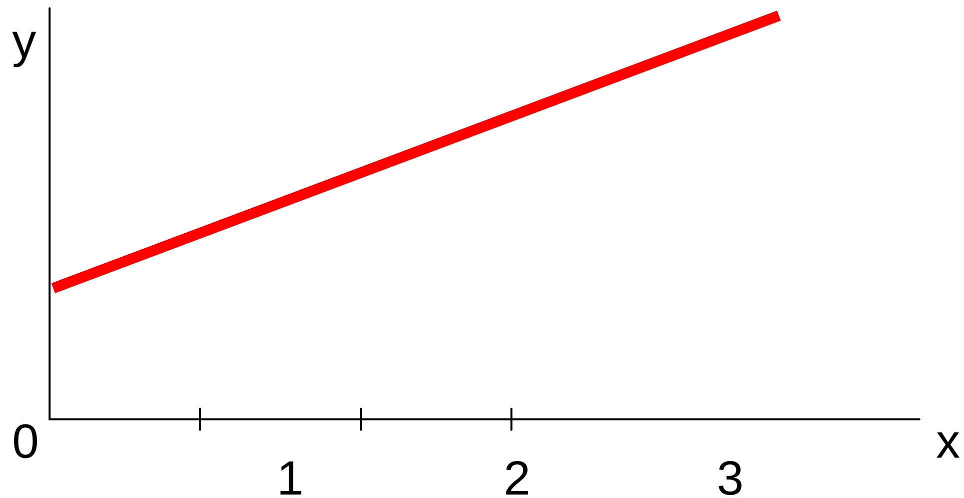
Prädiktorwert

vorhergesagter Wert

$$\boxed{\hat{y}_i = r_{xy} \cdot \frac{s_y}{s_x} \cdot \left(x_i - \bar{x}\right) + \bar{y}}$$

$$b_{yx} = r_{xy} \cdot \frac{s_y}{s_x}$$

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

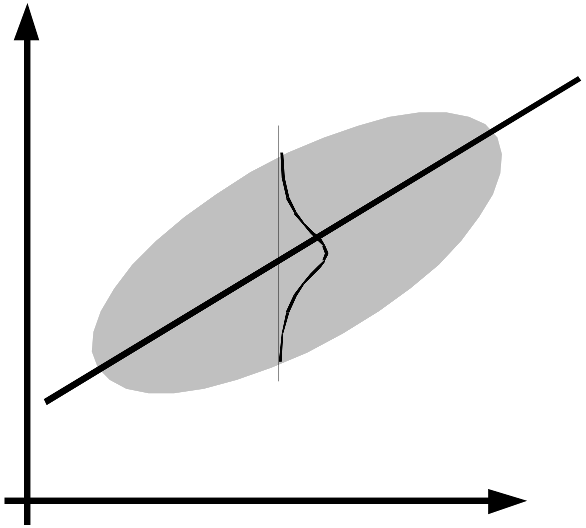


4. Vorraussetzungen

1.
2.
3.

4.
5.
6.

5. Güte der Vorhersage: Standardschätzfehler



=

•

•

Standardschätzfehler: Formeln

- Stichprobe:

$$s_{y.x} = s_y \cdot \sqrt{1 - r_{xy}^2}$$

- Populationsschätzung:

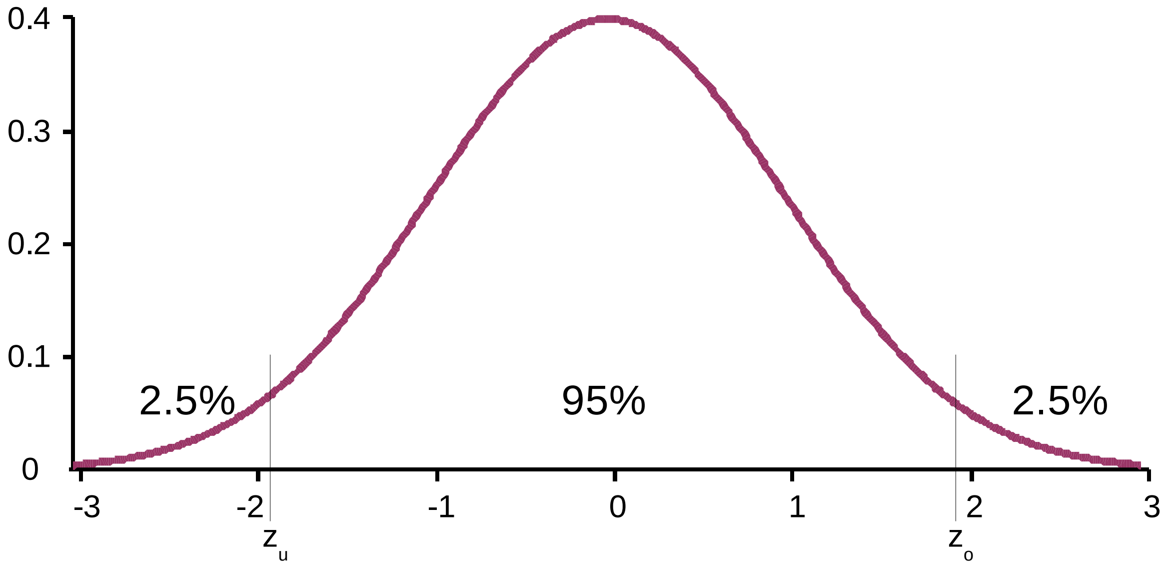
$$\boxed{\hat{\sigma}_{y.x} = \sqrt{\frac{N}{N-2}} \cdot s_{y.x}}$$

- nimmt ab (Schätzung wird genauer!):
  –
  –
  –

Konfidenzintervall

= Bereich, in dem ein wahrer Wert mit einer bestimmten Wahrscheinlichkeit liegt.

- Bei normalverteilten Variablen liegen 95% aller Werte in einem Bereich von Mittelwert ± 1.96 Standard-abweichungen.



Aus der z-Tabelle:

$z(p=0.025) = -1.96$

$z(p=0.975) = 1.96$

Konfidenzintervall

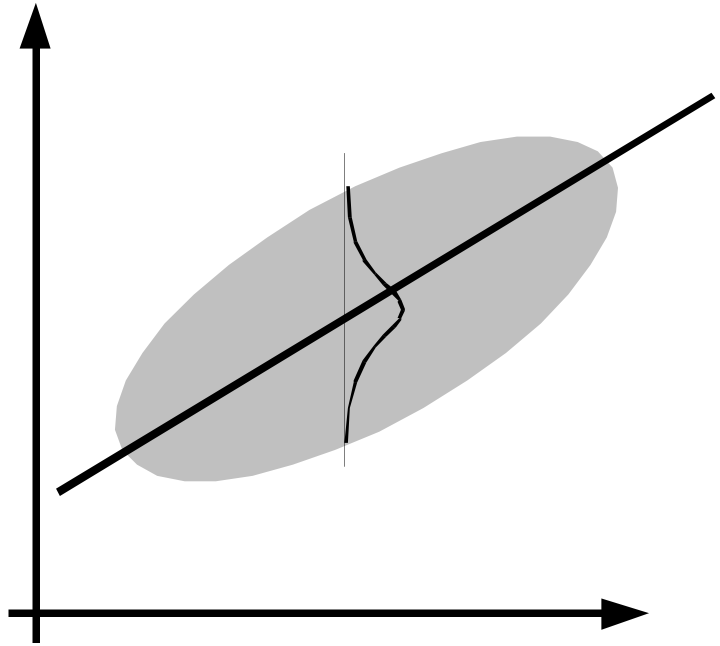
- Über den Standardschätzfehler, kann ein Konfidenzintervall berechnet werden, in dem mit bspw. 95%iger Wahrscheinlichkeit der wahre y-Wert liegt

- Stichprobe:

$$KI = \hat{y}_i \pm 1.96 \cdot \boxed{s_{y.x}}$$

- Populationsschätzung:

$$KI = \hat{y}_i \pm 1.96 \cdot \boxed{\hat{\sigma}_{y.x}}$$



6. Kreuzvalidierung: Überprüfung der externen Validität

(1) Berechnung der Regressionsgleichung R_1 anhand der Stichprobe S_1.

(2) Anwendung der Regressiongleichung R_1 auf die zweite Stichprobe S_2.

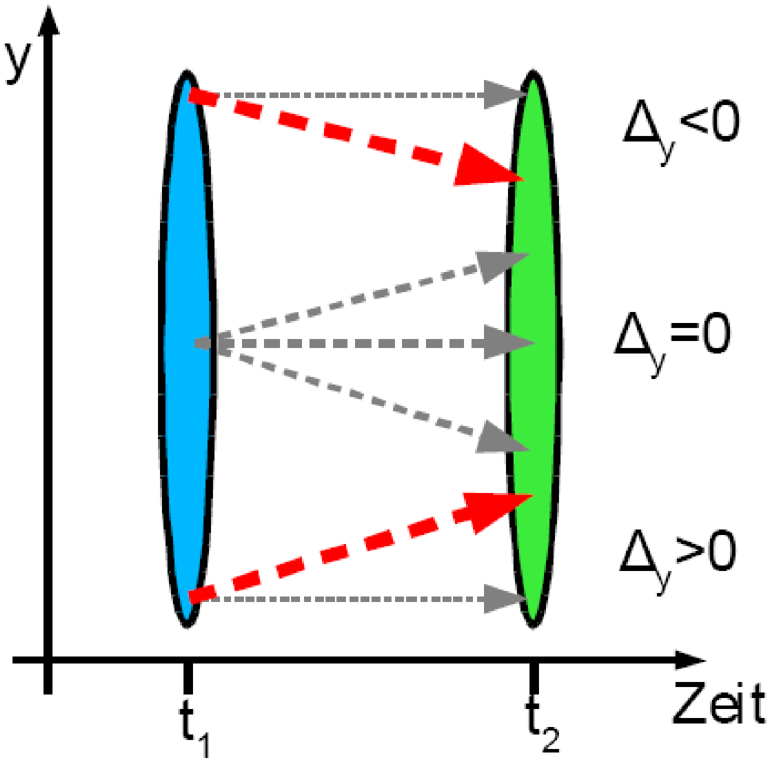
(3) Vergleich der vorhergesagten Kriteriumswerte mit den wahren Kriteriumswerten in S_2.

Berechnung der Regressionsgleichung R_2 anhand der Stichprobe S_2.

Anwendung der Regressiongleichung R_2 auf die Stichprobe S_1.

Vergleich der vorhergesagten Kriteriumswerte mit den wahren Kriteriumswerten in S_1.

7. Regression zur Mitte



- Problem, wenn bei wiederholter Messung die Personen der Stichprobe zu Beginn Extremwerte haben.
- Dann findet man statistisch einen Zusammenhang, der teilweise zufällig entstanden ist.

Übersicht über Abweichungen

- Varianz/ Standardabweichung
  - Maß für Streuung individueller Werte

$$\boxed{\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

- Standardfehler
  - Maß für Streuung der Stichprobenkennwerteverteilung (Mittelwert, Median)
  - Konfidenzintervall

$$\boxed{\widehat{\sigma}_{\overline{x}} = \sqrt{\frac{\widehat{\sigma}_x^{\,2}}{N}} = \frac{\widehat{\sigma}_x}{\sqrt{N}}}$$

- Standardschätzfehler
  - Maß für Streuung tatsächlicher y-Werte um Regressionsgerade
  - Konfidenzintervall

$$\boxed{\hat{\sigma}_{y.x} = \sqrt{\frac{n}{n-2}} \cdot s_{y.x}}$$