Motivation

- Wir betrachten die Zeit, die eine Festplatte zum Auslesen von Daten benötigt. Diese Lesezeit hängt ab von der Bewegung des Schreib–Lese–Kopfes, der Rotation der Platte und der Datenlänge.

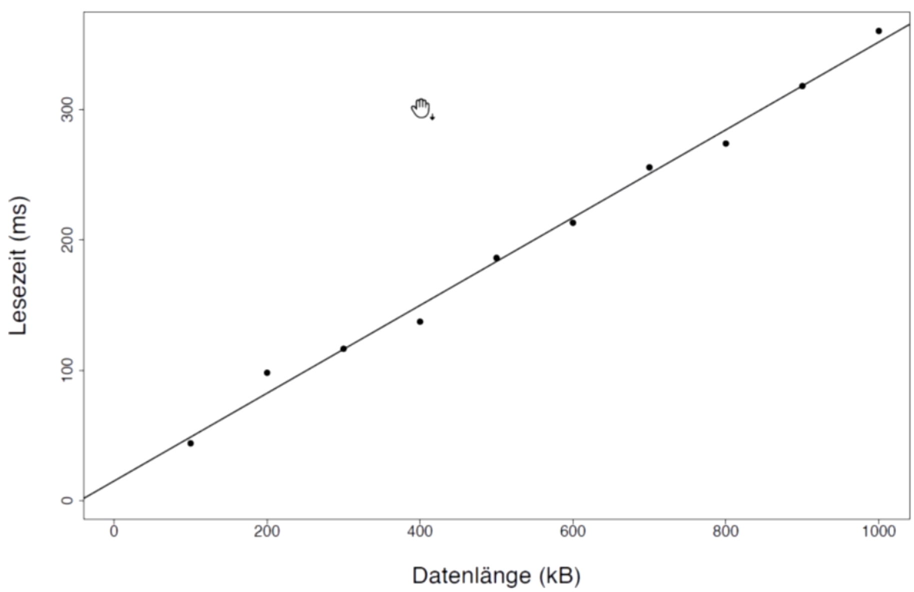
  Angenommen, die Lesezeit (ms) wird durch eine Zufallsvariable Y beschrieben. Für bestimmte Datenlängen x (kB) beobachten wir die folgenden Werte:

-

i	1	2	3	4	5
x_i	100	200	300	400	500
y_i	43.98	98.11	116.53	137.31	186.24

i	6	7	8	9	10
x_i	600	700	800	900	1000
y_i	213.17	255.67	273.97	318.05	360.31

-



Definition

- Für die Daten

$$(x_1, Y_1), \ldots, (x_n, Y_n)$$

ist das **einfache lineare Regressionsmodell** gegeben durch

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (i = 1, \ldots, n).$$

- Response Y

- [[Zufallsvariable]]
- Regressionsgerade

$$E(Y) = \beta_0 + \beta_1 x.$$

    *

- Prädiktor x
- Annahme
  - Fehler ϵ_i
    - i.i.d. mit $E(\epsilon_i) = 0$ und $\mathrm{Var}(\epsilon_i) = \sigma^2$.
    - normalverteilt $$Y_i \sim N(\mu_i, \sigma^2),$$
      - $\mu_i = E(Y_i) = \beta_0 + \beta_1 x_i$ für $i = 1, \ldots, n$.
  - Parameter $\beta_0, \beta_1, \sigma^2$
    - unbekannte Konstanten
    - β_0 Intercept,
    - β_1 Slope
- Ziel
  - Parameter anhand der Daten gutmöglichst beschreiben

Schätzer für β_0, β_1

- mittels Methode der kleinsten Quadrate
  - Dabei werden die Schätzer so gewählt, dass die Summe der quadrierten vertikalen Abstände zwischen den Y_i und der geschätzten Regressionsgerade minimal ist;

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2.$$

- Herleitung

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2 = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2 = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i.$$

–

– Normalgleichungen

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \stackrel{!}{=} 0,$$

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i \stackrel{!}{=} 0.$$

*

– Minimum

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \frac{s_{xY}^2}{s_{xx}^2},$$

*

$$s_{xY}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x}),$$

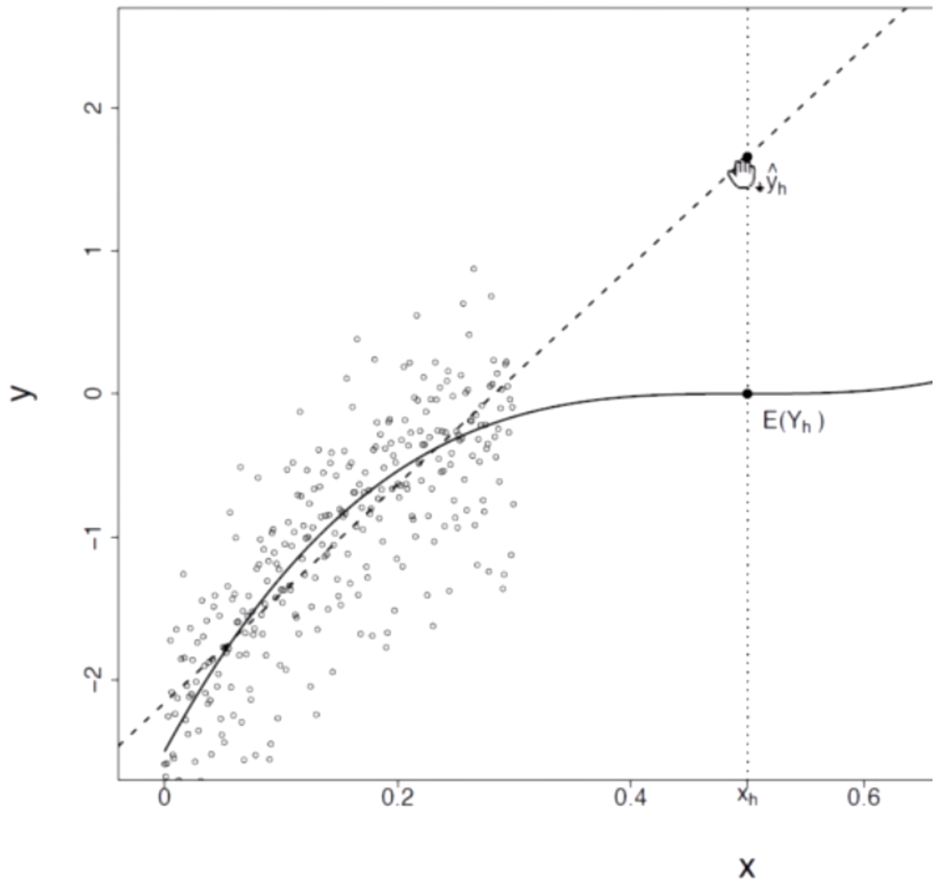
$$s_{xx}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

*

- geschätzte Regressionsgerade

$$\widehat{E(Y)} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

  -
  - sollte nicht für x außerhalb des beobachteten Bereichs benutzt werden
    - Extrapolation führt womöglich zu fehlerhaften Ergebnissen



    *

Beispiele

- Festplatte
  - siehe oben

	time	length
1	43.98	100
2	98.11	200
3	116.53	300
4	137.31	400
5	186.24	500
6	213.17	600
7	255.67	700
8	273.97	800
9	318.05	900
10	360.31	1000

–

– Berechnung

$$\bar{x} = 550 \qquad \text{und} \qquad \bar{Y} = 200.33,$$

*

$$(n-1)s_{xx}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{10}(x_i - 550)^2 = 825\,000,$$

$$(n-1)s_{xY}^2 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{10}(x_i - 550)(y_i - 200.33) = 277\,788.$$

*

$$\hat{\beta}_1 = \frac{s^2_{xY}}{s^2_{xx}} = \frac{(n-1)s^2_{xY}}{(n-1)s^2_{xx}} = \frac{277788}{825\,000} = 0.34,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 200.33 - 0.34 \cdot 550 = 15.14.$$

*

$$\widehat{E(Y)} = 15.14 + 0.34 \cdot x.$$

*