Fachhochschule
Dortmund
University of Applied Sciences and Arts

Data Science

09: Linear regression

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Recap

Given a random process with sample space Ω. A function X, mapping every possible result $\omega \in \Omega$ to a real number is called **random variable**:

$$X : \Omega \to \mathbb{R}, \omega \mapsto X(\omega) = x.$$

x is also known as the **realization of the random variable**.

Data Science

Recap Discrete random variable

A random variable X has a **discrete distribution** or the distribution of X is called **discrete distribution** if the sample space

$$\{x|X(\omega) = x, \omega \in \Omega_0\}$$

has countable many elements. The set

$$\Omega = \{x|P(X = x) > 0, x \in \mathbb{R}\}$$

is called **support** of X.

The support contains all possible realizations of X.

Data Science

Recap: Discrete distribution functions

Name	Density	Support	Expected value	Variance
Binomial				
$Bin(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,1,\ldots,n\}$	np	$np(1-p)$
Discrete uniform				
$DU(m)$	$\frac{1}{m}$	$\{0,1,\ldots,m\}$	$\frac{m+1}{2}$	$\frac{m^2-1}{12}$

Data Science

Recap

A random variable X has a **continuous distribution,** if there exists a function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) \geq 0$ for all $x \in \mathbb{R}$ such that

$$P(X \leq x) = F(X) = \int_{-\infty}^{x} f(t)dt$$

for all $x \in \mathbb{R}$ holds.
The function $f(x)$ is called **(probability) density** of X and $F(X)$ is called **distribution function**.

Note: $\{x|X(\omega) = x, \omega \in \Omega_0\}$ is a range or a union of ranges.

Data Science

Recap: Continuous distribution functions

Name	Density	Support	Expected value	Variance
Normal				
$N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\phi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mathbb{R}$	μ	σ^2
Standard normal				
$N(0,1)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$	$\mathbb{R}$	0	1
continuous uniform				
$Unif(a,b)$	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Data Science

Today

linear regression

Simple linear regression

Data Science

Simple linear regression

So far? Is there a (linear) correlation between two variables?

Examples

- Height of a father $\rightarrow$ Height of sons
- Temperature $\rightarrow$ Demand for winter shoes
- Body mass index (BMI) $\rightarrow$ Blood pressure
- Temperature, time $\rightarrow$ result of chemical processes
- Number of rooms, year of construction $\rightarrow$ house price

The **Pearson correlation coefficient** is a measure for the strength and direction of a linear correlation.

Data Science

Simple linear regression

Often, it is of interest to model the correlation between two variables! E.g.

- understanding why these values are correlated and how the influence is
- predicting the most likely values for unknown inputs

Modelling the linear correlation between an independent variable X and a dependent variable Y.

"For given x, what is the value of y"

Data Science

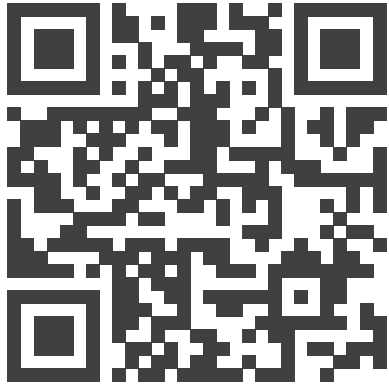
Simple linear regression

In the following we consider two examples:

- **Example from literature** [1]:
  - Cholesterol measurements for 24 patients with hyperlipoproteinemia
  - Assumption: Cholesterol level increases with age
  - Question: How well can the level of cholesterol be estimated from patients age?
  - Input: Age - Output: Cholesterol value
- **Own survey**:
  - Time from leaving home to the start of the lecture
  - Assumption: Starting time at home increases with distance to FH
  - Question: How well can we estimate the time from starting at home till start of the lecture from the distance to the FH?
  - Input: Distance - Output: time from starting at home till start of lecture

Data Science

Simple linear regression

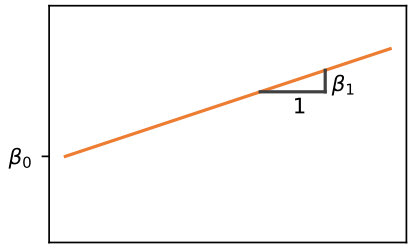


Data Science

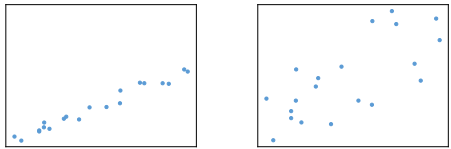
Simple linear regression

If two variables are perfectly linear correlated (Pearson coefficient 1), a value of y can be computed by

$$y = \beta_0 + \beta_1 x$$



Unfortunately, for observations this is nearly never the case ...



- **Problem:** For fixed x, Y is a random variable. Therefore, for given x, y can only be computed with an uncertainty.

Which value do we **expect** that the variable Y takes for given $X = x$ and how strong do the values **scatter** around?

Data Science

Simple linear regression

Let’s take a look

Simple linear regression

Model and parameter estimator

Data Science

Model and parameter estimator

We assume that for a given x the value y of Y is given by

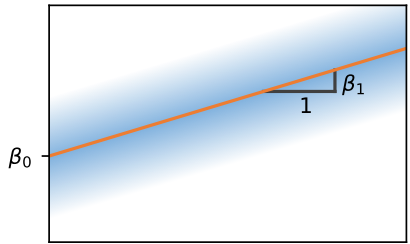
$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We assume that σ^2 is independent of x.

- β_0 is the intercept and β_1 the slope of the true regression line.
- X: **independent** variable also **predictor** or **explanatory variable** (not random)
- Y: **dependent** variable also **regressor** - for fixed x a random variable
- ϵ: random error term - for fixed x a random variable

Data Science

Model and parameter estimator



The expected value of Y is a function depending on x:

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

Data Science

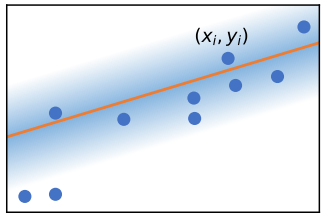
Model and parameter estimator

A sample is then given for values $x_1, \ldots, x_n$ with

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

or with values

$$(x_1, y_1), \ldots, (x_n, y_n).$$

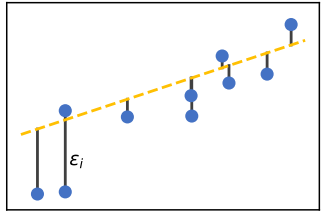


Data Science

Model and parameter estimator

The values of the sample (x_i, y_i) can be rewritten as:

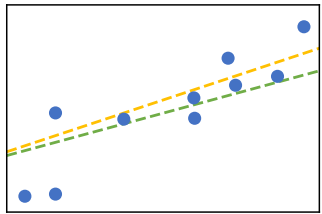
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



Data Science

Model and parameter estimator

Other way around: Only a sample is given - what are β_0 and β_1?



Which line is the ”best”? We need a predictor for the values β_0 and β_1!

Simple linear regression is a model that estimates the linear relationship between one independent and one dependent variable.

Data Science

Model and parameter estimator

How to find the ideal line?

Motivation (previous lecture): For an observation with values $x_1, \ldots, x_n$ the function $f(z) = \sum_i^n (x_i - z)^2$ takes its minimum in the mean value $\bar{x}$.

The observed mean value should be close to the expected value - **least squares method**!

Data Science

Model and parameter estimator

The least-squares method is a method to find the best fitting curve (line) by minimizing the squared distance between observations and line.

The **least-squares estimator** for fitting a linear curve is given by: Find $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}$ such that

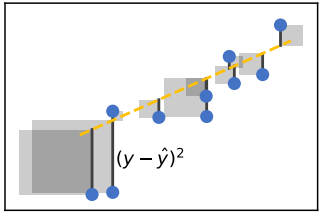
$$L(\hat{\beta}_0, \hat{\beta}_1) = \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} L(\beta_0, \beta_1)$$

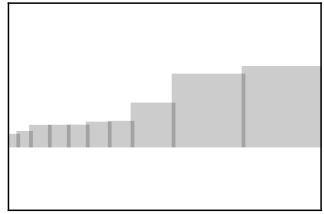
with

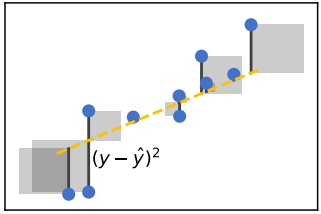
$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

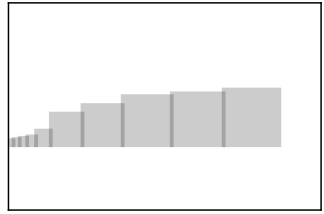
Data Science

Model and parameter estimator









Data Science

Model and parameter estimator

Let’s take a look

Problem: How to solve the minimization problem?

Often: Minimization problems can not be solved exactly - solution must be found numerically (e.g. via optimization). Luckily, for a simple linear regression, we can compute an analytical solution.

Least squares estimator must fulfill

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0 \text{ and } \left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0$$

Data Science

Model and parameter estimator

Task: How to compute $\hat{\beta}_0$ and $\hat{\beta}_1$?

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

Data Science

Model and parameter estimator

Task: How to compute $\hat{\beta}_0$ and $\hat{\beta}_1$?

$$\begin{aligned}
\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta}_1 x_i = 0 \\
&\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\beta}_0 \\
&\Leftrightarrow \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}
\end{aligned}$$

Data Science

Model and parameter estimator

Task: How to compute $\hat{\beta}_0$ and $\hat{\beta}_1$?

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0$$

Data Science

Model and parameter estimator

Task: How to compute $\hat{\beta}_0$ and $\hat{\beta}_1$?

$$\begin{aligned}
\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0 \\
&\Leftrightarrow \sum_{i=1}^{n}\left(y_i - (\overline{y} - \hat{\beta}_1\overline{x}) - \hat{\beta}_1 x_i\right)x_i = 0 \\
&\Leftrightarrow \sum_{i=1}^{n}(y_i - \overline{y})x_i = \hat{\beta}_1\sum_{i=1}^{n}(\overline{x} - x_i)x_i \\
&\Leftrightarrow \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x} + \overline{x}) = \hat{\beta}_1\sum_{i=1}^{n}(\overline{x} - x_i)(x_i - \overline{x} + \overline{x}) \\
&\Leftrightarrow \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) + \overline{x}\sum_{i=1}^{n}(y_i - \overline{y}) = \hat{\beta}_1\sum_{i=1}^{n}(\overline{x} - x_i)^2 + \overline{x}\hat{\beta}_1\sum_{i=1}^{n}(\overline{x} - x_i)
\end{aligned}$$

Data Science

Model and parameter estimator

$$\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) + \overline{x}\sum_{i=1}^{n}(y_i - \overline{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(\overline{x} - x_i)^2 + \overline{x}\hat{\beta}_1 \sum_{i=1}^{n}(\overline{x} - x_i)$$

Data Science

Model and parameter estimator

$$\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})+\overline{x}\sum_{i=1}^{n}(y_i-\overline{y})=\hat{\beta}_1\sum_{i=1}^{n}(\overline{x}-x_i)^2+\overline{x}\hat{\beta}_1\sum_{i=1}^{n}(\overline{x}-x_i)$$

$$\Leftrightarrow\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})=\hat{\beta}_1\sum_{i=1}^{n}(\overline{x}-x_i)^2$$

$$\Leftrightarrow\hat{\beta}_1=\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(\overline{x}-x_i)^2}$$

Data Science

Model and parameter estimator

The linear regression line for Y given X with observations $(x_i, y_i) \in \mathbb{R}^2$ for $i = 1, \ldots, n$ is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

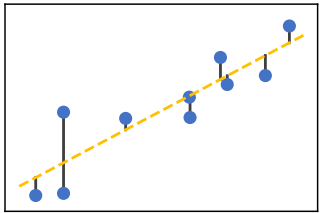
with

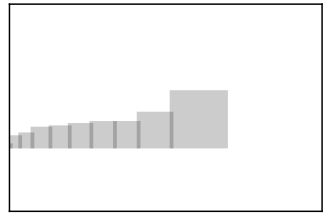
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (\overline{x} - x_i)^2} \text{ and } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- If ϵ is normally distributed, then are also β_0 and β_1 normally distributed.

Data Science

Model and parameter estimator





Data Science

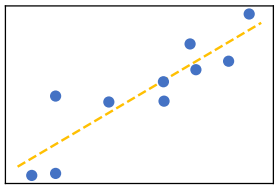
Model and parameter estimator

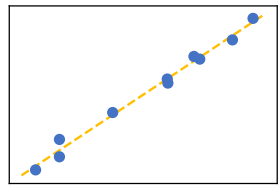
Let’s take a look

Data Science

Model and parameter estimator

The least square estimator gives the best fitting linear line for the given data.





How large is the variance of the underlying data?

Data Science

Model and parameter estimator

The **residual sum of squares (RSS)** for a simple linear regression is given by

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The expected value for the residual sum of squares depends on the variance, i.e.

$$E(RSS) = (n-2)\sigma^2$$

An estimator for the variance is given by

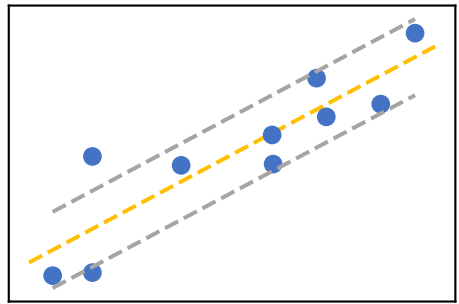
$$\hat{\sigma}^2 = \frac{RSS}{n-2}.$$

Why $n-2$? Number of parameter in regression (β_0 and β_1).

Data Science

Model and parameter estimator

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1.87	10.93	10.33	0.60	0.36
4.75	15.84	14.95	0.89	0.79
⋮	13.21	14.27	-1.07	1.14
3.01	10.98	12.15	-1.17	1.36
3.54	14.17	13.01	1.16	1.35



$$0.36 + 0.79 + \cdots + 1.35 = 15.99 \quad \Rightarrow \quad \sigma^2 = \frac{15.99}{10 - 2} \approx 2$$

Data Science

Model and parameter estimator

Let’s take a look

Simple linear regression

Residual analysis

Data Science

Residual analysis

Simple linear regression and **residual sum of squares** are estimators for β_0, β_1 and σ. But, only if the formulated requirements are fulfilled:

Reminder

We assume that for a given x the value y of Y is given by

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We assume that σ^2 is independent of x.

- Are these requirements fulfilled?

We can not prove it, but we can collect arguments!

Data Science

Residual analysis

From the **model assumption** $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ a **simple linear regression** gives $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The difference between observation and regression value is called **residuum**:

$$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- Every observation pair (x_i, y_i) fulfills

$$y_i = \hat{\beta}_i + \hat{\beta}_1 x_i + e_i$$

Note that e_i and ϵ_i are **not** the same!

- ϵ_i is a random variable with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- e_i is a value (error estimate of the i^{th} observation)

Data Science

Residual analysis

The investigation of the fit of the regression line using a residual plot is called **residual analysis**.

- Residual plot is a specialized scatter plot
- Residual is plotted in relation to x_i or y_i values
- It is common to plot the residual values

$$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i \text{ for } i \in \{1, \ldots, n\}$$

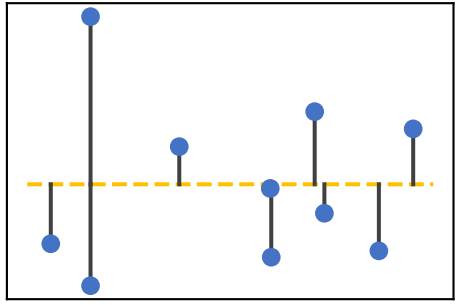
or standardized residual values

$$\hat{d}_i = \frac{y_i - \hat{y}_i}{\sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}} \text{ for } i \in \{1, \ldots, n\}$$

- Zero line corresponds to the fitted line

Data Science

Residual analysis



Data Science

Residual analysis

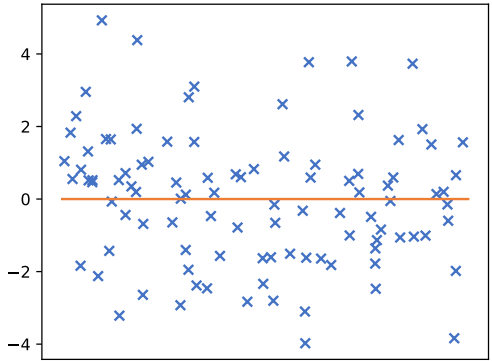
Assuming that all assumptions are fulfilled, then the following must hold or nearly hold:

- The zero-line ($y = 0$) is located in the center of the image
- The residues show the same scattering
- There are no further structures visible

If all conditions are fulfilled, we can assume that also our assumptions are fulfilled.

Data Science

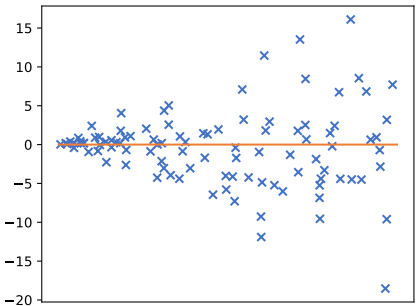
Residual analysis

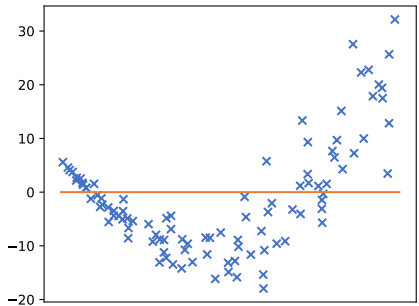


Residual plot fulfilling all conditions formulated before.

Data Science

Residual analysis





Residual plot **not** fulfilling all conditions formulated before.

Data Science

Residual analysis

Let’s take a look

Simple linear regression

Determinacy measure

Data Science

Determinacy measure

Given a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n$$

with least square estimator $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, then there holds

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

Scatter decomposition

$$TSS = ESS + RSS$$

- TSS = Total sum of squares
- ESS = Error sum of squares
- RSS = Regression sum of squares

The value

$$R^2 = \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

is called **coefficient of determination**.

There holds

- $0 \leq R^2 \leq 1$
- $R^2 = 1 - \frac{ESS}{TSS}$
- $R^2 = 1$ if and only if $e_i = 0$ for all $i = 1, \ldots, n$: optimal fit, i.e. all observations lie on the regression line.
- $R^2 = 0$ if and only if $\hat{y}_i = \overline{y}$ for all $i = 1, \ldots, n$

Data Science

Determinacy measure

- For a simple linear regression there holds

$$R^2 = r_{XY}^2$$

A large value of R^2 means that the model fits the data well - but when is a value large?

- A few rules of thumb:
  - For a technical application: Good fit if $R > 0.8$.
  - For a social-scientific application: Good fit if $R > 0.5$

Data Science

Determinacy measure

Let’s take a look

Summary & Outlook

Data Science

Summary & Outlook: Summary

- You understand what a simple linear regression is and can describe the parameters
- You are able to perform a simple linear regression
- You can estimate the variance
- You can perform a residual analysis and interpret the results
- You can compute and interpret the R^2 value and interpret the results

Statistical tests

References

Data Science

Summary & Outlook: Endnotes

[1] Sachs, L., Hedderich, J. (2009), Angewandte Statistik - Methodensammlung mit R, 13. Auflage, Springer, Heidelberg

Data Science

Summary & Outlook: Acknowledgement

Parts of the lecture base on the lecture ”Statistics” (FH Dortmund)
by
Prof. Dr. Sonja Kuhnt and Prof. Dr. Nadja Bauer.