Optimization and Data Science

Lecture 7: Regression

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

1. Regression
    - Overview
    - Linear vs. Nonlinear Regression
    - Optimality Conditions for a Linear Regression Problem: The Normal Equations
    - Examples, Disadvantages and Possible Problems
    - Interpretation of Regression Results

Contents

1. Regression
   - Overview
   - Linear vs. Nonlinear Regression
   - Optimality Conditions for a Linear Regression Problem: The Normal Equations
   - Examples, Disadvantages and Possible Problems
   - Interpretation of Regression Results

Regression

- What is it?

  Construction of an approximative, **reduced-order** model for given data, e.g., a linear or quadratic model to describe a more complex dataset

- Why are we studying this?

  One of the most popular ways to analyze data

  Also used to predict future behavior (**to be used with care!)**

- How does it work?

  Defining the model structure (e.g., linear model)

  Then optimization of the model parameters to obtain best fit to data

- What if we can use it?

  Find "structure" of or "behind" data

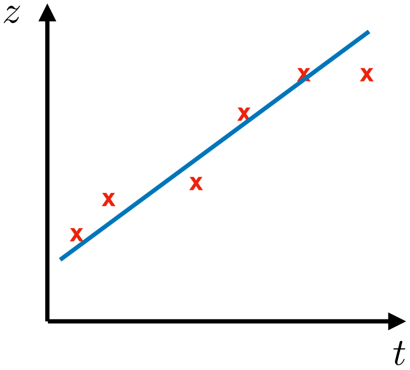
  Detect basic behavior of or trends in data

  Make predictions (**to be used with care!)**

Example from 1st lecture: Data-fitting by linear regression

- Given: data points

$$(t_k, z_k)_{k=1,\dots,m},\ t_k, z_k \in \mathbb{R}.$$



- Observation: approx. linear dependency
- Task: Detect parameters of this dependency
- Mathematical task: Find affine-linear function

$$y(t) = at + b$$

that satisfies (at least approximately)

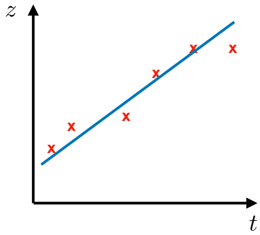
$$y(t_k) = at_k + b \approx z_k, \quad k = 1, \dots, m.$$

- Exact equality not possible for $m > 2$
- $\rightsquigarrow$ minimize distance between points and function (optimization problem)

Linear regression: The optimization problem behind

- Minimize (squared) distance between function and data

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{k=1}^{m} (at_k + b - z_k)^2.$$



- Rewrite with $x := (a, b), A \in \mathbb{R}^{m\times 2}$ for $k = 1, \dots, m$:

$$y(t_k) = at_k + b = t_k a + 1 \cdot b = A_{k1}x_1 + A_{k2}x_2 = \sum_{j=1}^{2} A_{kj}x_j = (Ax)_k,$$

with

$$A = \begin{pmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{pmatrix} \in \mathbb{R}^{m\times 2}, \quad x = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2.$$

Linear regression: The optimization problem behind

- Rewritten with $x := (a, b), A \in \mathbb{R}^{m\times 2}$

$$\sum_{k=1}^{m}(at_k + b - z_k)^2 = \sum_{k=1}^{m}(Ax - z)_k^2 = \|Ax - z\|_2^2$$

with

$$A = \begin{pmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{pmatrix} \in \mathbb{R}^{m\times 2}, \quad x = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2.$$

$\rightsquigarrow$ Minimize (squared) distance between function and data

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{k=1}^{m}(at_k + b - z_k)^2 \iff \min_{x\in\mathbb{R}^2} \|Ax - z\|_2^2.$$

Linear regression problems

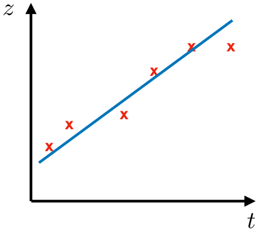
Definition

Let $z \in \mathbb{R}^m$ be data and $y = Ax \in \mathbb{R}^m$ a given linear model, i.e., y depends linearly on some parameters $x \in \mathbb{R}^n$. The problem to find

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} \|Ax - z\|_2^2$$

is called **linear regression problem** or **linear least-squares problem**.

- Linear regression: function (model) y depends *linearly on the parameters* x.
- It is *not important* that the function y (in the example above) was *a linear function with respect to* t.
- The function y may also be called a **reduced-order model**.



Contents

1 Regression

- Overview
- Linear vs. Nonlinear Regression
- Optimality Conditions for a Linear Regression Problem: The Normal Equations
- Examples, Disadvantages and Possible Problems
- Interpretation of Regression Results

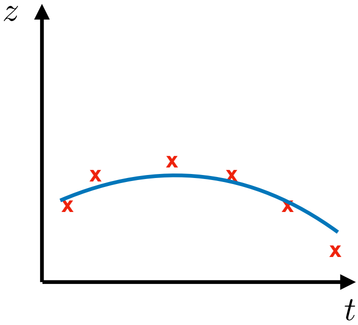
Difference: linear vs. nonlinear Regression

- As above: datapoints

$$(t_k, z_k)_{k=1,\dots,m},\ t_k, z_k \in \mathbb{R}.$$

- But now: find quadratic function such that

$$y(t_k) = at_k^2 + bt_k + c \approx z_k, \quad k = 1, \dots, m.$$



$\rightsquigarrow$ again optimization problem:

$$\min_{(a,b,c)\in\mathbb{R}^3} \sum_{k=1}^{m} (at_k^2 + bt_k + c - z_k)^2 \Longleftrightarrow \min_{x\in\mathbb{R}^3} \|Ax - z\|_2^2.$$

with

$$A = \begin{pmatrix} t_1^2 & t_1 & 1 \\ \vdots & \vdots & \vdots \\ t_m^2 & t_m & 1 \end{pmatrix} \in \mathbb{R}^{m\times 3}, \quad x = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \in \mathbb{R}^3.$$

Difference: linear vs. nonlinear Regression

- Analogously (**linear** w.r.t. the parameters):

$$\text{Polynomial:} \quad y(t) = \sum_{k=0}^{n} a_j t^j, \quad x = (a_j)_{j=0}^{n}$$

$$\text{Trigonometric polynomial:} \quad y(t) = \sum_{j=1}^{n} a_j \sin(jt), \quad x = (a_j)_{j=1}^{n}.$$

- But:

$$y(t) = ae^{bt}, \quad x = (a, b) \in \mathbb{R}^2,$$

cannot be written as

$$y(t_k) = (Ax)_k, \quad k = 1, \dots, m,$$

- Function y is linear w.r.t. parameter a, but nonlinear w.r.t. parameter b.
- $\rightsquigarrow$ **Nonlinear** regression problem.

Linear and nonlinear regression problems

Definition (as above)

Let $z \in \mathbb{R}^m$ be data and $y = Ax \in \mathbb{R}^m$ a given linear model, i.e., y depends linearly on some parameters $x \in \mathbb{R}^n$. The problem to find

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} \|Ax - z\|_2^2$$

is called **linear regression problem** or **linear least-squares problem**.

Definition

Let $z \in \mathbb{R}^m$ be data and $y = F(x)$ a given model, where $F : \mathbb{R}^n \to \mathbb{R}^m$ is nonlinear. The problem to find

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} \|F(x) - z\|_2^2$$

is called **nonlinear regression problem** or **nonlinear least-squares problem**.

Contents

1. Regression
   - Overview
   - Linear vs. Nonlinear Regression
   - Optimality Conditions for a Linear Regression Problem: The Normal Equations
   - Examples, Disadvantages and Possible Problems
   - Interpretation of Regression Results

Optimality conditions for a linear regression problem

- We apply the first and second order optimality conditions on the linear regression problem

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{where } f(x) = \|Ax - z\|_2^2 = \sum_{i=1}^{m} (Ax - z)_i^{\,2} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} x_j - z_i \right)^2$$

- Partial derivatives:

$$\begin{aligned}
\frac{\partial f}{\partial x_k}(x) &= \sum_{i=1}^{m} \frac{\partial}{\partial x_k} \left(\sum_{j=1}^{n} A_{ij} x_j - z_i \right)^2 = \sum_{i=1}^{m} 2 \left(\sum_{j=1}^{n} A_{ij} x_j - z_i \right) A_{ik} \\
&= 2 \sum_{i=1}^{m} A_{ik} (Ax - z)_i = 2 \sum_{i=1}^{m} \left(A^\top \right)_{ki} (Ax - z)_i \\
&= 2 \left(A^\top (Ax - z) \right)_k, \qquad k = 1, \dots, n.
\end{aligned}$$

1st order necessary optimality condition: Normal equations

- Partial derivatives:

$$\frac{\partial f}{\partial x_k}(x) = 2\left(A^\top(Ax - z)\right)_k, \quad k = 1, \ldots, n.$$

- 1st order condition:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_k}(x)\right)_{k=1}^{n} = 2A^\top(Ax - z) = 0 \quad \Longleftrightarrow \quad A^\top Ax = A^\top z.$$

- This system of equations is called **normal equation(s)**.

$$\boxed{A^\top \in \mathbb{R}^{n\times m}}\ \boxed{A \in \mathbb{R}^{m\times n}} = \boxed{\begin{array}{c} A^\top A \\ \in \mathbb{R}^{n\times n}\end{array}}$$

Geometric interpretation: normal equations

- Normal equations:

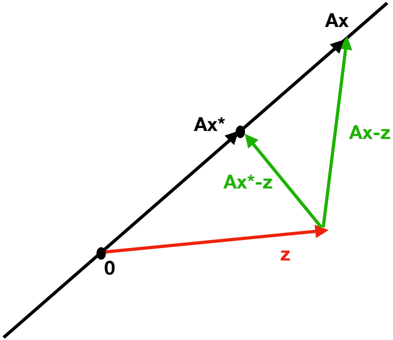
$$A^\top (Ax^* - z) = 0.$$

- $\{Ax \in \mathbb{R}^m : x \in \mathbb{R}^n\}$ is linear subspace of $\mathbb{R}^m$.
- x^* is a minimizer of $\|Ax - z\|_2$, if

$$Ax \perp (Ax^* - z) \quad \forall x \in \mathbb{R}^n,$$

- Especially for $x = e_i$ (unit coordinate vectors): We have $Ae_i = A_{*i}$ and

$$\begin{aligned} & A_{*i} \perp (Ax^* - z) && \forall i = 1, \ldots, n \\ \Leftrightarrow \quad & A_{*i}^\top (Ax^* - z) = 0 && \forall i = 1, \ldots, n \\ \Leftrightarrow \quad & A^\top (Ax^* - z) = 0. \end{aligned}$$



2nd order conditions for a linear regression problem

- Partial derivatives:

$$\frac{\partial f}{\partial x_k}(x) = 2\sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}x_j - z_i\right) A_{ik}, \quad k = 1,\dots,n.$$

- 2nd partial derivatives:

$$\frac{\partial^2 f}{\partial x_\ell \partial x_k}(x) = \frac{\partial f}{\partial x_\ell}\left(2\sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}x_j - z_i\right) A_{ik}\right) = 2\sum_{j=1}^{n} \underbrace{A_{i\ell}}_{=\left(A^\top\right)_{\ell i}} A_{ik} = 2\left(A^\top A\right)_{\ell k},$$

$$k,\ell = 1,\dots,n.$$

$\rightsquigarrow$ Hessian matrix:

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_\ell \partial x_k}(x)\right)_{\ell,k=1}^{n} = 2A^\top A.$$

2nd order optimality conditions

- Hessian matrix:
$$\nabla^2 f(x) = 2A^\top A \text{ for all } x \in \mathbb{R}^n,$$
- ... is constant and positive semi-definite:
$$x^\top 2A^\top Ax = 2(Ax)^\top Ax = 2\|Ax\|_2^2 \geq 0 \text{ for all } x \in \mathbb{R}^n.$$
- Here we used:
$$(Ax)^\top = x^\top A^\top \text{ and } \|x\|_2^2 = \sum_{i=1}^n x_i^2 = x^\top x.$$
- The Hessian is positive definite, if A has full rank.

We obtain:

Theorem

A solution $x \in \mathbb{R}^n$ to the normal equation $A^\top Ax = A^\top z$ is a solution of the linear regression problem with data $z \in \mathbb{R}^m$ and matrix $A \in \mathbb{R}^{m\times n}$. If A has full rank, the solution is unique.

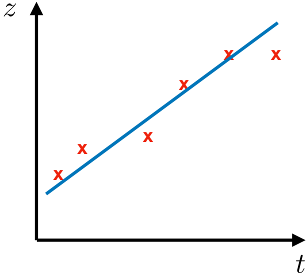
Contents

1. Regression
   - Overview
   - Linear vs. Nonlinear Regression
   - Optimality Conditions for a Linear Regression Problem: The Normal Equations
   - Examples, Disadvantages and Possible Problems
   - Interpretation of Regression Results

Example: Regression line

- Model function:

$$y(t) = at + b$$



- Matrix:

$$A = \begin{pmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{pmatrix}$$

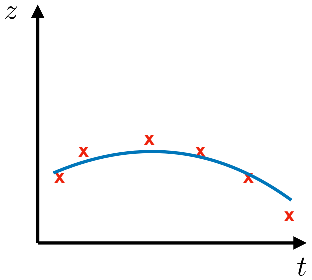
- ... gives:

$$A^\top A = \begin{pmatrix} t_1 & \cdots & t_m \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^m t_k^2 & \sum_{k=1}^m t_k \\ \sum_{k=1}^m t_k & m \end{pmatrix}$$

Example: Regression with polynomial of higher order

- Model function (polynomial of order n):

$$y(t) = \sum_{j=0}^{n} a_j t^j$$



- Matrix:

$$A = \begin{pmatrix} t_1^n & \cdots & 1 \\ \vdots & & \vdots \\ t_m^n & \cdots & 1 \end{pmatrix}$$

- ... gives:

$$A^\top A = \begin{pmatrix} t_1^n & \cdots & t_m^n \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} t_1^n & \cdots & 1 \\ \vdots & & \vdots \\ t_m^n & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{m} t_k^{2n} & \cdots & \sum_{k=1}^{m} t_k \\ \vdots & & \vdots \\ \sum_{k=1}^{m} t_k & \cdots & m \end{pmatrix}$$

Disadvantages and problems that might occur in regression

- Effort to compute $A^\top A$. How many operations are necessary?
- Entries in the matrix $A^\top A$ may have wide spread in magnitude:

$$\sum_{k=1}^{m} t_k^{2n} \gg m.$$

- Example: Polynomial of degree $n = 3$, date points $t = 1, \ldots, m = 100$:

$$A^\top A = \begin{pmatrix} 1.4791e+13 & 1.7171e+11 & 2.0503e+09 & 2.5502e+07 \\ 1.7171e+11 & 2.0503e+09 & 2.5502e+07 & 3.3835e+05 \\ 2.0503e+09 & 2.5502e+07 & 3.3835e+05 & 5.0500e+03 \\ 2.5502e+07 & 3.3835e+05 & 5.0500e+03 & 1.0000e+02 \end{pmatrix}$$

- $\rightsquigarrow$ Solution of the linear system will be sensitive to errors.
- Nearly linear dependent rows in matrix $\rightsquigarrow$ matrix "nearly" singular $\rightsquigarrow$ result (solution of linear system) may be inaccurate.

Different norms for approximation?

- We could use a different norm to solve

$$\min_{x\in\mathbb{R}^n} \|Ax - z\|$$

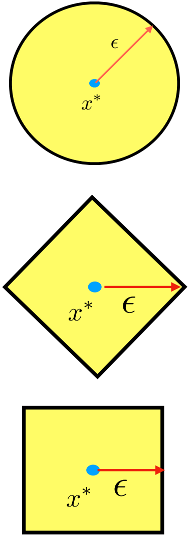
- But note: These two norms

$$\|x\|_1 := \sum_{i=1}^n |x_i|$$
$$\|x\|_\infty := \max_{i=1,\dots,n} |x_i|.$$

are not differentiable.

$\rightsquigarrow$ cannot apply 1st+2nd order conditions

$\rightsquigarrow$ cannot compute the solution by normal equations.

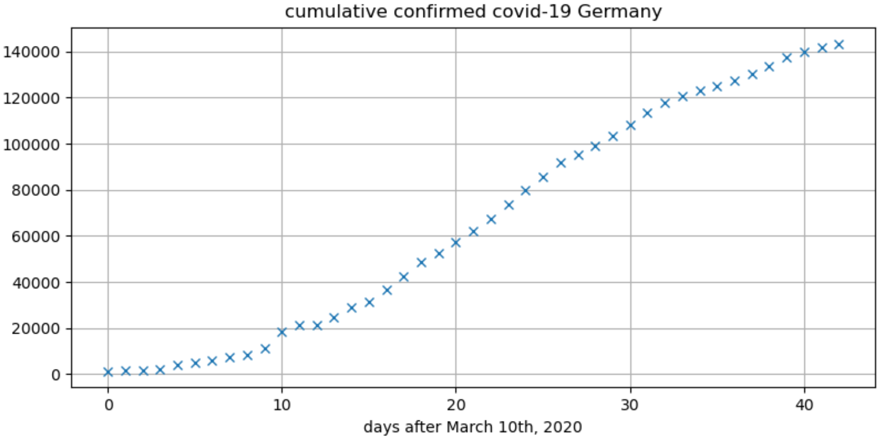


Contents

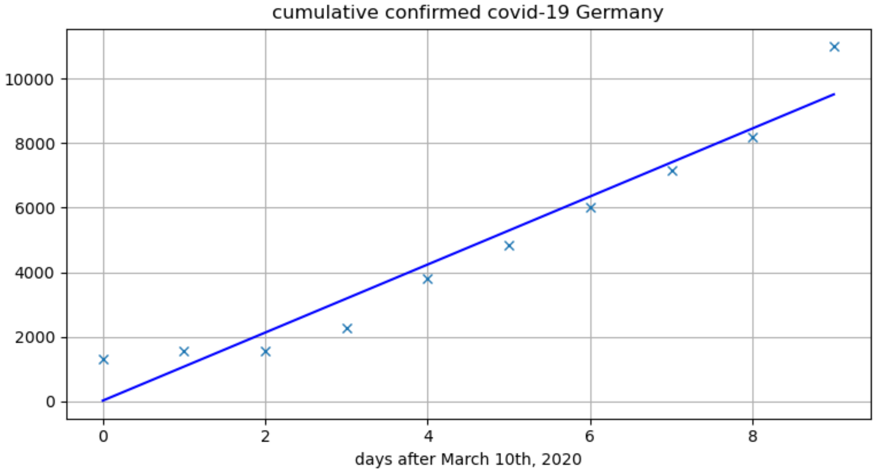
1. Regression
   - Overview
   - Linear vs. Nonlinear Regression
   - Optimality Conditions for a Linear Regression Problem: The Normal Equations
   - Examples, Disadvantages and Possible Problems
   - Interpretation of Regression Results

Regression results to understand data or the underlying process

Again: data points $(t_k, z_k)_{k=1,\dots,m}$, $t_k, z_k \in \mathbb{R}$. Data: WHO `https://covid19.who.int`.

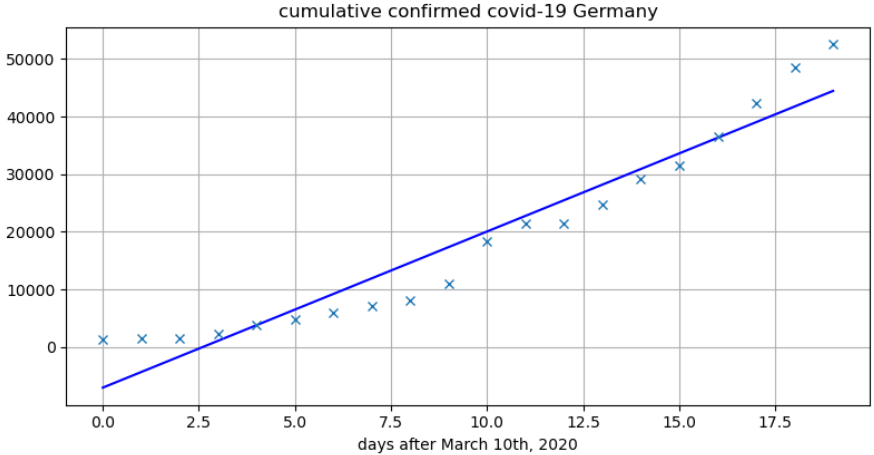


Building different regression lines



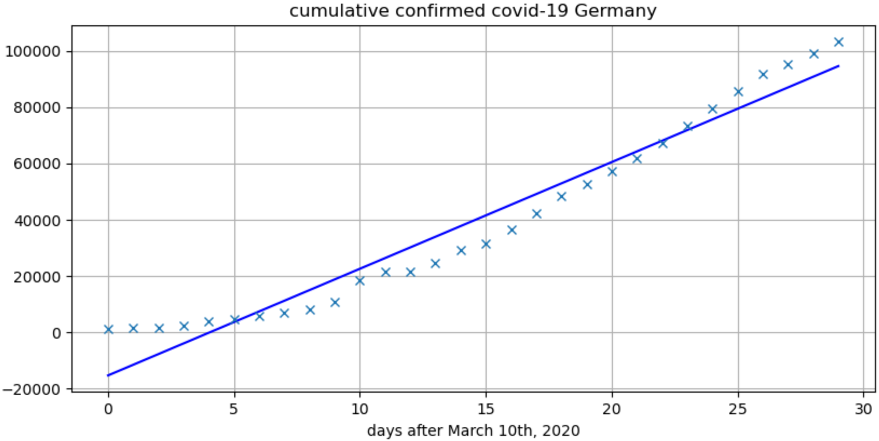
slope: $a \approx 1054$

Building different regression lines



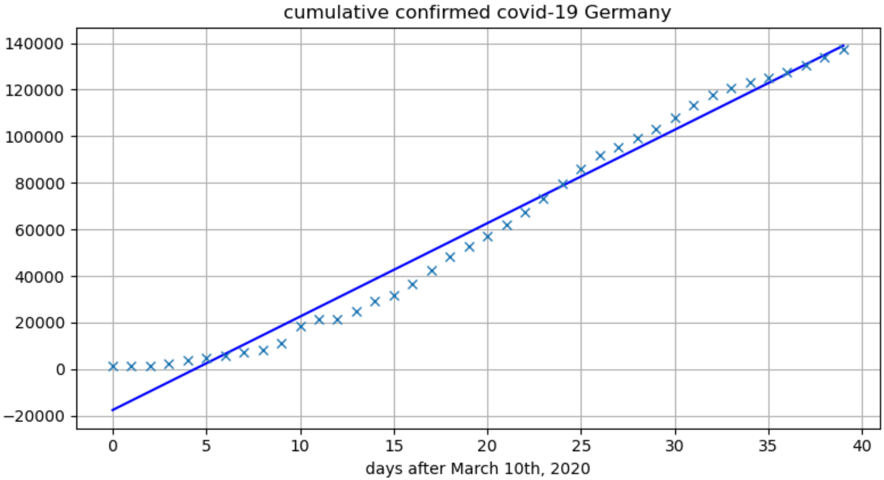
slope: $a \approx 2710$

Building different regression lines



slope: $a \approx 3790$

Building regression line (40 data-points)



slope: $a \approx 4011$

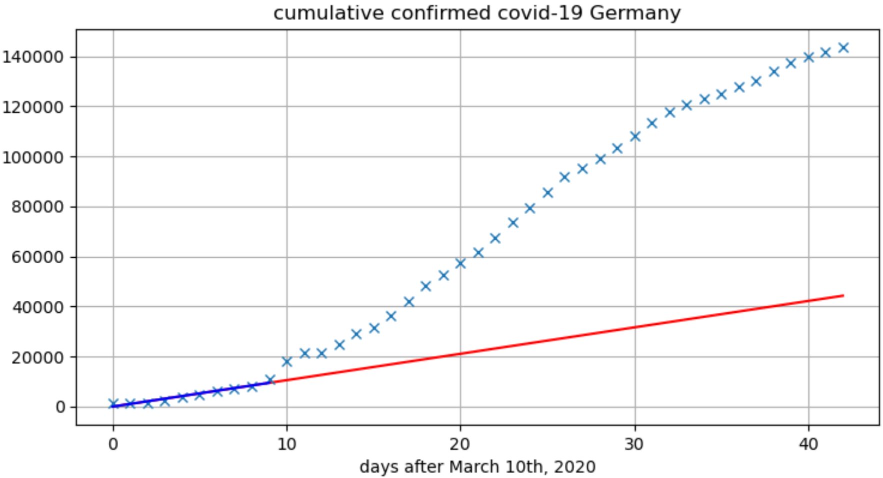
Regression results as predictions

- Regression lines look ok for analysis ...
- They provide a data-based model.
- Often: data points are measurements from the past, t is time:

$$(t_k, z_k)_{k=1,\dots,m},\, t_k, z_k \in \mathbb{R}.$$

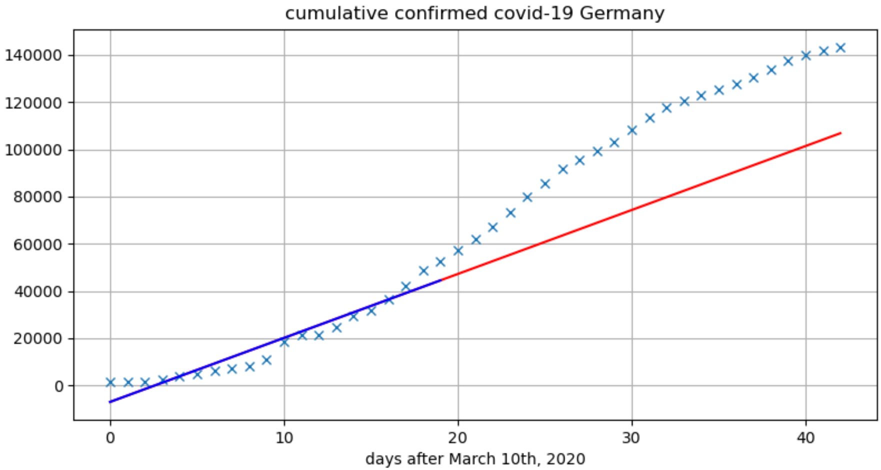
- Idea might be: take regression line (i.e., data-based model) as projection for the future ..
- ... meaning: for points outside the used (time) interval where the data t are coming form.
- This might be misleading, since projection into the future (= outside the range of the data) is purely speculative.

Prediction by data-based model: underestimation (first 10 data-points)



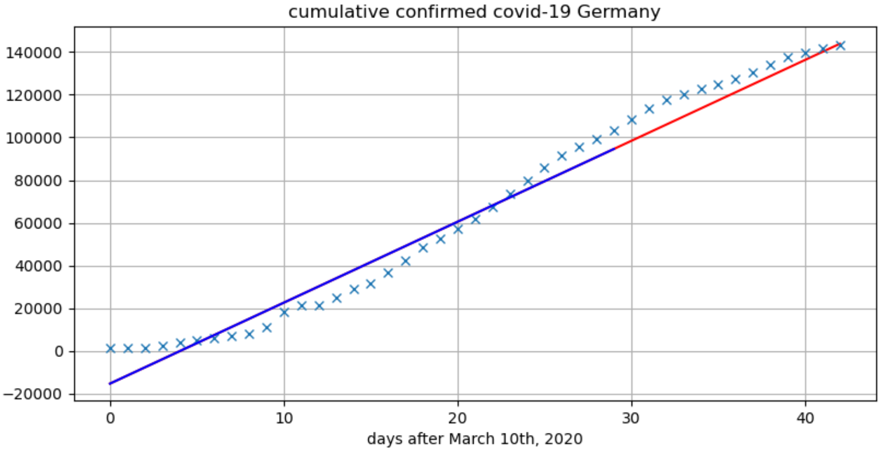
slope: $a \approx 1054$

Prediction by data-based model: underestimation (first 20 data-points)



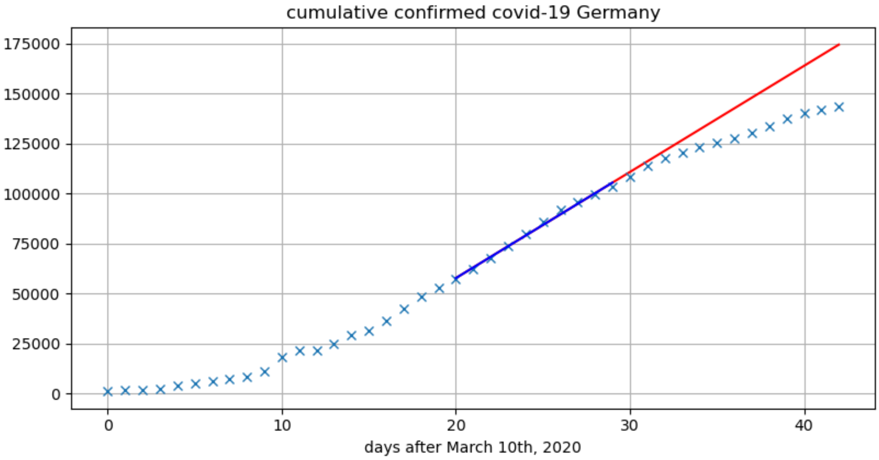
slope: $a \approx 2710$

Prediction by data-based model (first 30 data-points)



slope: $a \approx 3790$

Prediction: overestimation (10 data-points since 20th March)



slope: $a \approx 5305$

What is important?

- Regression is an approximation of given data by lower-order functions/models.
- Easiest case: regression line.
- Optimization problem behind is a least-squares problem for the model parameters.
- Therein, we minimize the squared Euklidean norm of the pointwise distance between data and model.
- First and second order optimality conditions lead to a system of linear equations, called normal equations. Its solution is the solution to the linear regression problem.
- Regression can be used to interpret data.
- For prediction, it has to be used with care, since the model is based on given data only.