Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Mathe III

Ralf Herbrich

Regression

Überblick

Überblick

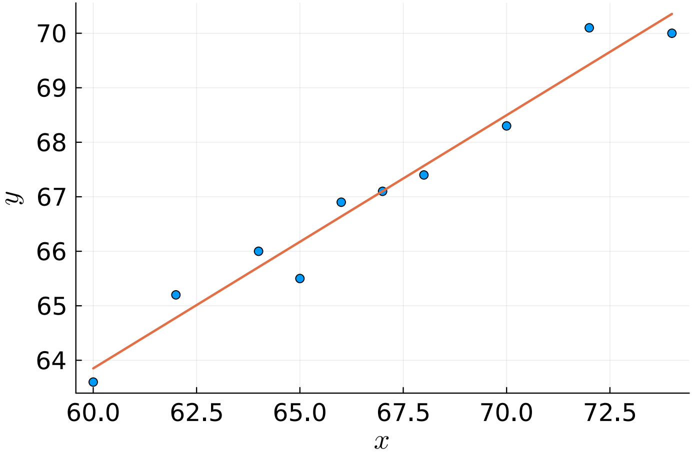
Motivation: Hypothesentests für Lineare Regression

- **Beispiel (Vererbung von Größen).** Im Jahre 1886 untersuchte Sir Francis Galton, ob es einen genetischen Zusammenhang zwischen den Größen von Eltern und Kindern gibt. Dazu erstellte er einen Datensatz aus der Körpergröße von 930 erwachsenen Kindern und deren 205 Eltern. Karl Pearson wiederholte das Experiment mit 10 Vätern und deren erwachsenen Söhnen. Unterstützen die Daten die Hypothese, dass Kinder **im Mittel** kleiner sind als ihre Eltern?
- **Maximum *Likelihood* Regression:**

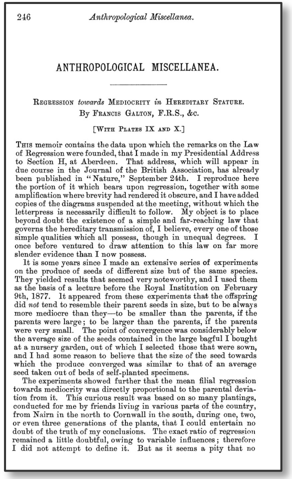
```
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    35.9768       0.4646
```

Initiale Bestätigung. **Aber**: Wie sicher können wir uns aufgrund der Stichprobe sein?



Sir Francis Galton
(1822 – 1911)



246 *Anthropological Miscellanea.*

ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

Wiederholung: Hypothesentests & Verteilungen

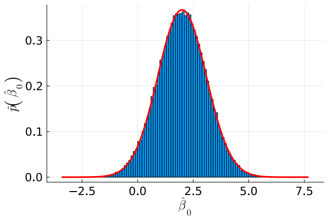
- **Design eines Hypothesentests: Konstruktion in 4 Schritten**
  1. Festlegen eines parametrischen Modells $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$
  2. Formulierung von Null- und Alternativhypothese
  3. Wahl eines Irrtumsniveaus α
  4. Konstruktion einer Ablehnungsregion (*rejection region*) unter Nullhypothese
- Verteilungsannahme unter Nullhypothese hilft bei Konstruktion der Ablehnungsregion!
- **Satz (Verteilung von $\widehat{\beta_0}$, $\widehat{\beta_1}$ und SSR).** Im linearen Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit unabhängigem Fehlern $Z \sim \mathcal{N}(0, \sigma^2)$ gilt für ein Stichprobe der Größe n

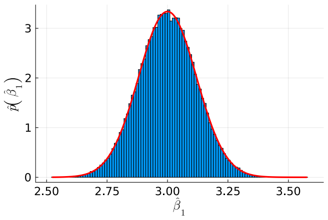
$$\widehat{\beta_0} \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \cdot \sum_{i=1}^n x_i^2}{n \cdot \left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)}\right) \qquad \widehat{\beta_1} \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}\right)$$

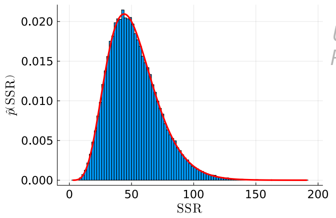
SSR

$$\frac{\boxed{\sum_{i=1}^n \left(y_i - \left(\widehat{\beta_1} \cdot x_i + \widehat{\beta_0}\right)\right)^2}}{\sigma^2} \sim \chi^2(n-2)$$

Mathe III



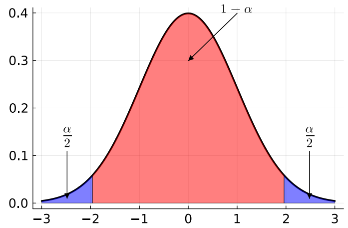




Überblick

1. Konzept der Linearen Regression
2. Maximum *Likelihood* Lineare Regression
   - Methode der Kleinsten Quadrate
3. Verteilungen der Maximum *Likelihood* Schätzer
4. Hypothesentests für Regressionsparameter
   - **t-Test für Lineare Regression**
5. Analyse der Residuen
6. Exkurs: Multiple Regression

t-Tests mit einer Stichprobe



- **Daten**: Ein Datensatz $(\boldsymbol{x}, \boldsymbol{y})$ von n Beobachtungen $(x_i, y_i) \in \mathbb{R}^2$ und ein Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit unabhängigem Fehlern $Z \sim \mathcal{N}(0, \sigma^2)$ bei unbekannten Parameter β_0, β_1 und unbekannter Varianz σ^2.
- **Nullhypothese** $H_0\colon \beta_1 = 0$
- **Teststatistik**: Wir betrachten den MLE $\widehat{\beta_1}$, und unter H_0 gilt

$$\widehat{\beta_1} \sim \mathcal{N}\left(0, \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}\right) \Leftrightarrow \frac{\widehat{\beta_1}}{\sigma} \cdot \sqrt{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2} \sim \mathcal{N}(0,1)$$

  Außerdem gilt

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2(n-2) \quad \Rightarrow \sqrt{\frac{(n-2) \cdot \left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)}{\sum_{i=1}^n \left(y_i - \left(\widehat{\beta_1} \cdot x_i + \widehat{\beta_0}\right)\right)^2}} \cdot \widehat{\beta_1} \sim t(n-2)$$

- **Annahmeregion**:

$$\left|\sqrt{\frac{(n-2) \cdot \left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)}{\sum_{i=1}^n \left(y_i - \left(\widehat{\beta_1} \cdot x_i + \widehat{\beta_0}\right)\right)^2}} \cdot \widehat{\beta_1}\right| \le \tau_{1-\frac{\alpha}{2}}$$

t-Tests mit einer Stichprobe in R

- R vereinfacht den t-Test, indem es alle relevanten Werte für einen Hypothesentest der Regressionsparameter ausrechnet (mittels **summary**)!

```
x = c(  60,   62, 64,   65,   66,   67,   68,   70,   72, 74)
y = c(63.6, 65.2, 66, 65.5, 66.9, 67.1, 67.4, 68.3, 70.1, 70)
father_son = lm(y ~ x)

summary(father_son)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6738 -0.2374 -0.0852  0.2835  0.6742

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.97681    2.20760   16.30 2.02e-07 ***
x            0.46457    0.03298   14.08 6.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4321 on 8 degrees of freedom
```

Kommandoaufruf von einer Zusammenfassung einer Regressionsanalyse

Deskriptive Statistik der Residuen

$$\hat{\sigma}_{\widehat{\beta_0}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \left(\widehat{\beta_1}\cdot x_i + \widehat{\beta_0}\right)\right)^2}{(n-2)\cdot n\cdot\left(\sum_{i=1}^{n} x_i^2 - n(\boldsymbol{x})^2\right)}\cdot\sum_{i=1}^{n} x_i}$$

$\widehat{\beta_0}$

$\widehat{\beta_1}$

$$\hat{\sigma}_{\widehat{\beta_1}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \left(\widehat{\beta_1}\cdot x_i + \widehat{\beta_0}\right)\right)^2}{(n-2)\cdot\left(\sum_{i=1}^{n} x_i^2 - n(\boldsymbol{x})^2\right)}}$$

$\dfrac{\widehat{\beta_0}}{\hat{\sigma}_{\widehat{\beta_0}}}$

p-Wert für Test auf $\widehat{\beta_0} = 0$

p-Wert für Test auf $\widehat{\beta_1} = 0$

$\dfrac{\widehat{\beta_1}}{\hat{\sigma}_{\widehat{\beta_1}}}$

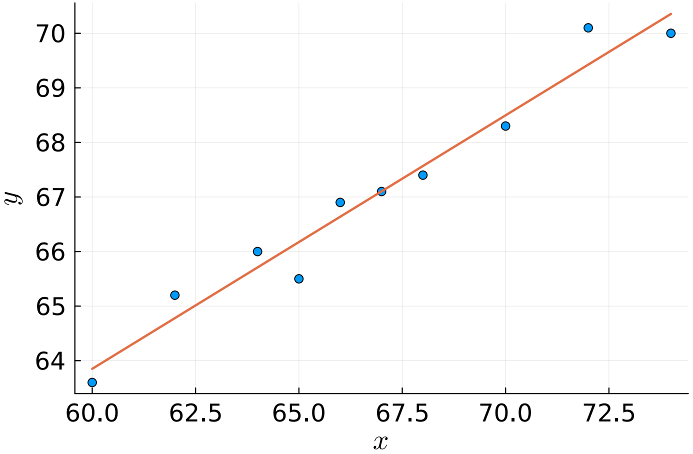
Beispiel: t-Test für Lineare Regression

- **Beispiel (Vererbung von Größen)**. Im Jahre 1886 untersuchte Sir Francis Galton, ob es einen genetischen Zusammenhang zwischen den Größen von Eltern und Kindern gibt. Dazu erstellte er einen Datensatz aus der Körpergröße von 930 erwachsenen Kindern und deren 205 Eltern. Karl Pearson wiederholte das Experiment mit 10 Vätern und deren erwachsenen Söhnen. Unterstützen die Daten die Hypothese, dass Kinder **im Mittel** kleiner sind als ihre Eltern?
- **Hypothesentest**:

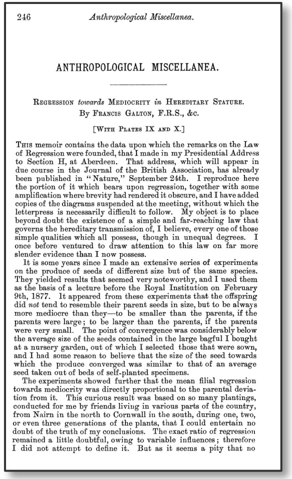
```
summary(father_son)

             Estimate Std. Error t value
x             0.46457    0.03298   14.08
```

  - Hypothese H_0: $\beta_1 \geq 1$
  - Teststatistik: $\frac{\widehat{\beta}_1 - 1}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{0.46457 - 1}{0.03298} = -16.23$
  - p-Wert ($\sim t(8)$): $1.045569 \cdot 10^{-7}$ $\Rightarrow$ "H_1"



Sir Francis Galton
(1822 – 1911)



246 *Anthropological Miscellanea.*

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.
By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

Bemerkungen: t-Test für Lineare Regression

- **Bemerkungen (t-Test für Lineare Regression)**
  - Das Prinzip des t-Tests ist direkt überführt worden, nachdem die Verteilungen der Regressionsparameter sowohl normalverteilt und unabhängig χ^2-verteilt waren.
  - Der Test ist auch anwendbar, wenn mehr als eine unabhängige Variable existiert („multiple“ Regression).
  - Der Test kann auch benutzt werden, um festzustellen, ob unabhängige Variablen überhaupt statistisch signifikanten Einfluss auf die Vorhersage haben.
  - Dieser Test wird auch im Maschinellen Lernen eingesetzt, um Modellparameter zu „*prunen*“.
  - Man kann den Test auch anpassen, um Hypothesentests für einzelne Vorhersagen für zukünftigen x_i zu machen (und Konfidenzintervalle an diesen Vorhersagen angeben).
  - Mehrschichtige neuronale Netzwerke („*deep neural networks*“) sind nichts anderes, als hintereinander angewandte Regressionen mit einer komponentenweisen, festen, nicht-linearen Transferfunktion!

Überblick

1. Konzept der Linearen Regression
2. Maximum *Likelihood* Lineare Regression
   - Methode der Kleinsten Quadrate
3. Verteilungen der Maximum *Likelihood* Schätzer
4. Hypothesentests für Regressionsparameter
   - t-Test für Lineare Regression
5. **Analyse der Residuen**
6. Exkurs: Multiple Regression

Gütetests für Modelle mit Linearer Regression

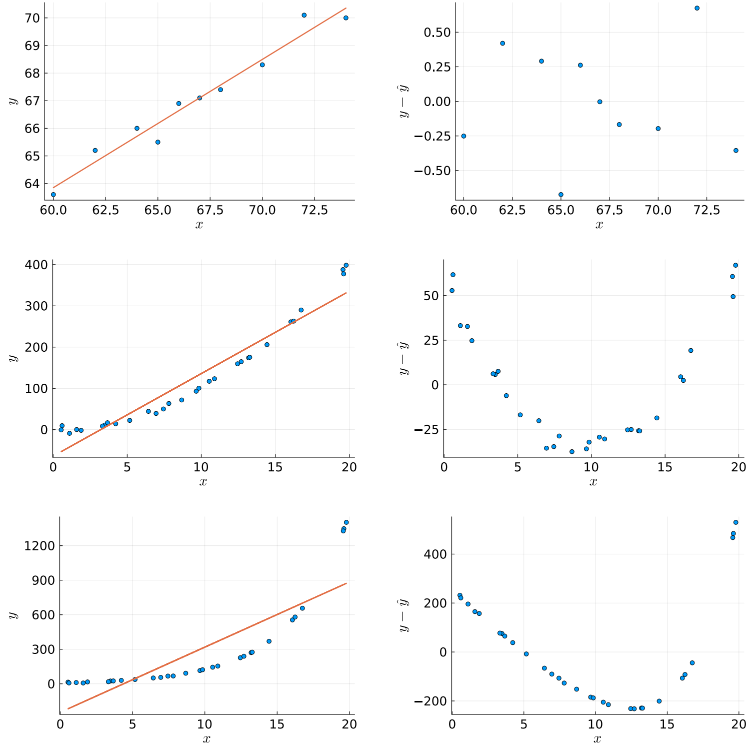
- **Frage**: Wie können wir **visuell** überprüfen, ob ein lineares Modell eine gute Modellannahme war?
- **Beobachtung**: Im linearen Regressionsmodell haben wir angenommen, dass

$$Y_i = \beta_1 \cdot x_i + \beta_0 + Z_i$$

  wobei die $Z_i \sim \mathcal{N}(0, \sigma^2)$ und paarweise unabhängig.
- **Idee**: Wenn die Modellannahme richtig ist, dann müssten die Z_i paarweise unabhängig sein (und **nicht** von x_i abhängen)!
  - Für einen Datensatz von n Datenpunkten (x_i, y_i) bekommen wir n Realisierungen von Z

$$z_i = y_i - \left(\widehat{\beta_1} \cdot x_i + \widehat{\beta_0}\right)$$

  - Wenn wir diese **Residuen** gegen x_i plotten, sollte eine unkorrelierte Punktwolke entstehen



Transformation von unabhängigen Variablen

- Wenn die Analyse der Residuen vermuten lässt, dass nichtlineare Zusammenhänge bestehen, wie z.B.

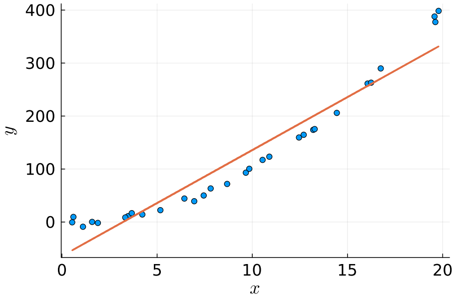
  $$Y_i = \beta_1 \cdot x_i^p + \beta_0 + Z_i$$

  dann können wir eine Regression mit den transformierten **unabhängigen** Variablen

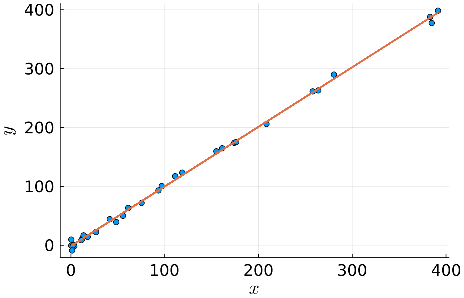
  $$\tilde{x}_i = x_i^p$$

  versuchen und erhalten unter Umständen ein besseres Modell!

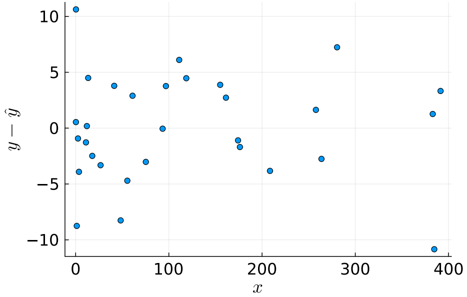
$Y_i = \beta_1 \cdot x_i + \beta_0 + Z_i$



$Y_i = \beta_1 \cdot \tilde{x}_i + \beta_0 + Z_i$



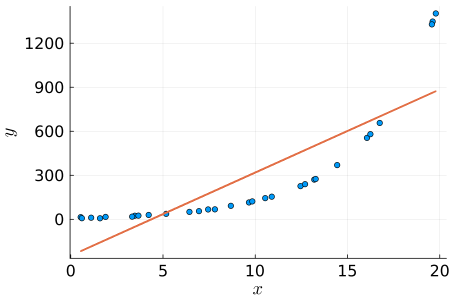
$y_i - \left(\widehat{\beta_1} \cdot \tilde{x}_i + \widehat{\beta_0}\right)$



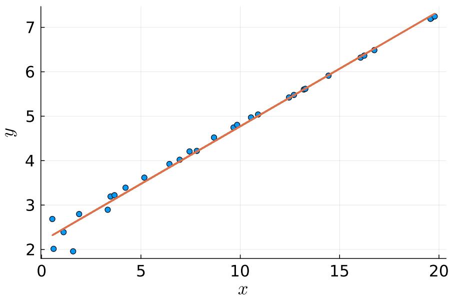
Transformation von abhängigen Variablen

- Wenn die Analyse der Residuen vermuten lässt, dass
$$Y_i = c \cdot \exp(\beta_1 \cdot x_i) + Z_i$$
dann können wir eine Regression mit den transformierten **abhängigen** Variablen
$$\tilde{y}_i = \log(y_i)$$
versuchen und erhalten unter Umständen ein besseres Modell!

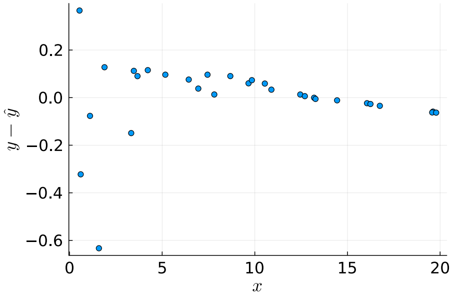
$Y_i = c \cdot \exp(\beta_1 \cdot x_i) + Z_i$



$\tilde{Y}_i = \beta_1 \cdot x_i + \beta_0 + Z_i$



$\tilde{y}_i - (\beta_1 \cdot x_i + \beta_0)$



Überblick

Exkurs: Lineare Modelle mit Basisfunktionen

- **Definition (Lineare Modelle mit Basisfunktionen).** Gegeben einen Eingaberaum $\mathcal{X}$ und D Basisfunktionen $\phi_i : \mathcal{X} \to \mathbb{R}$, ist ein lineares Modell mit Basisfunktionen eine Funktion

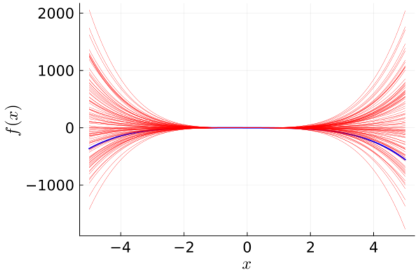
$$f(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 \cdot \phi_1(x) + \beta_2 \cdot \phi_2(x) + \cdots + \beta_D \cdot \phi_D(x)$$

  - **Bemerkung.** Das Modell ist linear in $\boldsymbol{\beta}$, nicht linear in $x \in \mathcal{X}$!

- **Beispiel.** 4 Basisfunktionen ($D = 4$) und 100 zufällige Parametervektoren $\boldsymbol{\beta}$.

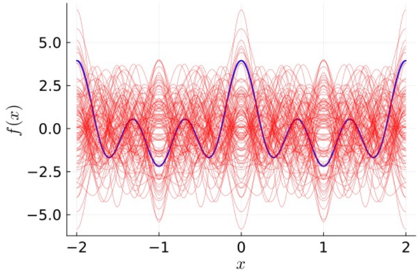
Polynomial Basis

$$\phi_j(x) = x^j$$



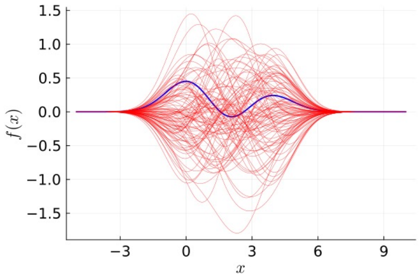
Fourier Basis

$$\phi_j(x) = \cos(\pi j \cdot x)$$



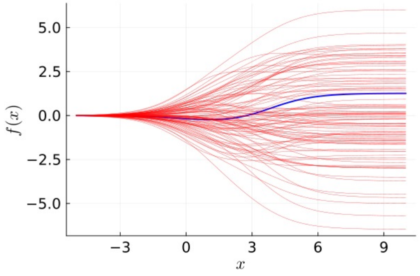
Gaussian Basis

$$\phi_j(x) = \mathcal{N}(x; j, 1)$$



Sigmoid Basis

$$\phi_j(x) = \frac{\exp(x - j)}{1 + \exp(x - j)}$$



Exkurs: Maximum *Likelihood* und Kleinste Quadrate

- **Normalverteiltes Fehlermodell**. Jede Beobachtung Y_i wird durch die Addition einer normalverteilten Zufallsvariable auf $f(x_i;\boldsymbol{\beta})$ erzeugt

$$Y_i = f(x_i;\boldsymbol{\beta}) + Z_i\,, \qquad Z_i \sim \mathcal{N}(0,\sigma^2)$$

- **Maximum *Likelihood***. Der Maximum *Likelihood* Schätzer von $\boldsymbol{\beta}$ ist gegeben durch

$$\boldsymbol{\beta}_{\mathrm{ML}} := \operatorname{argmax}_{\boldsymbol{\beta}} \prod_i p(y_i) = \operatorname{argmax}_{\boldsymbol{\beta}} \sum_i \log\left(\mathcal{N}(y_i; f(x_i;\boldsymbol{\beta}),\sigma^2)\right)$$

  - **Bemerkung**. Der Logarithmus ist eine strikt monotone Funktion, aus Produkten werden Summen und die Optimierung ist numerisch stabiler (Warum?)

- **Satz (Kleinste Quadrate)**. Der Minimierer der Funktion $\sum_i \left(y_i - f(x_i;\boldsymbol{\beta})\right)^2$ ist $\boldsymbol{\beta}_{\mathrm{ML}}$.
  - **Beweis**. Wenn wir die Dichte der Normalverteilung einsetzen, erhalten wir für die Summanden

$$\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i - f(x_i;\boldsymbol{\beta})\right)^2}{2\sigma^2}\right)\right) = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\cdot\left(y_i - f(x_i;\boldsymbol{\beta})\right)^2$$

Der einzige Teil der von $\boldsymbol{\beta}$ abhängt

Exkurs: Multiple Regression

- Wenn wir die $n \times D$ Matrix $\mathbf{\Phi}$ definieren als $\mathbf{\Phi}_{ij} = \phi_j(x_i)$ und alle Beobachtungen in einem n-dimensionalen Vektor $\boldsymbol{y}$ zusammenfassen, dann gilt in Matrixschreibweise

$$\boldsymbol{\beta}_{\mathrm{MLE}} \coloneqq \mathrm{argmin}_{\boldsymbol{\beta}} \|\mathbf{\Phi}\boldsymbol{\beta} - \boldsymbol{y}\|^2$$

$= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\boldsymbol{\beta} - 2\boldsymbol{y}^T\mathbf{\Phi}\boldsymbol{\beta} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = g(\boldsymbol{\beta})$

- Mit Hilfe der ersten Ableitung von $g(\boldsymbol{\beta})$ können wir zeigen, dass

$$\left.\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{MLE}}} = 2\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\boldsymbol{\beta}_{\mathrm{ML}} - 2\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$$

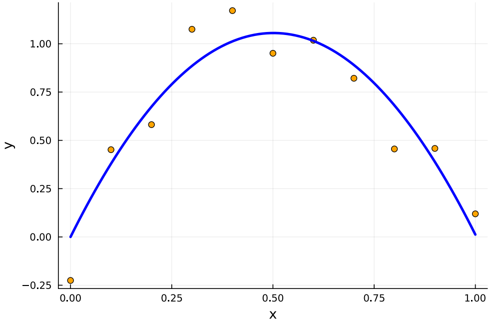
$$\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)\boldsymbol{\beta}_{\mathrm{MLE}} = \mathbf{\Phi}^{\mathrm{T}}\boldsymbol{y}$$

$$\boldsymbol{\beta}_{\mathrm{MLE}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{y}$$

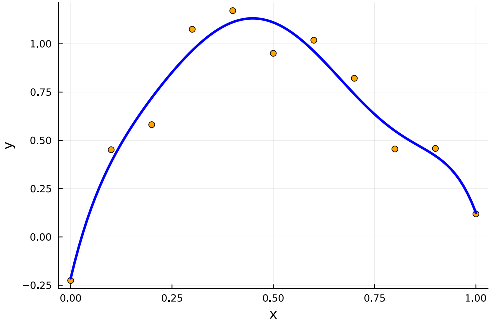
- **Bemerkungen**.
  - Die lineare Regression ist ein Spezialfall für $\phi_0(x) = 1$ und $\phi_1(x) = x$.
  - Der Aufwand der Berechnung ist $\max\left(\mathcal{O}(n \cdot D^2), \mathcal{O}(D^3)\right)$.
  - Die zweite Ableitung ist $2\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ welche per Definition positive-semidefinit ist; daher ist dies ein Minimum!

Multiple Regression: Polynomielle Basisfunktionen

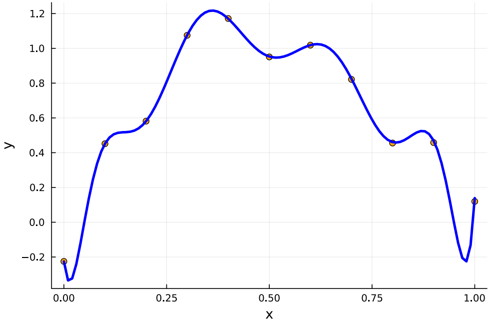
$$f(x;\boldsymbol{\beta}) = \beta_1 x + \beta_2 x^2$$



$$f(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6$$



$$f(x;\boldsymbol{\beta}) = \sum_{i=0}^{10} \beta_i \cdot x^i$$



Viel Spaß bis zur nächsten Vorlesung!