HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Mathe III

Ralf Herbrich

Regression

Überblick

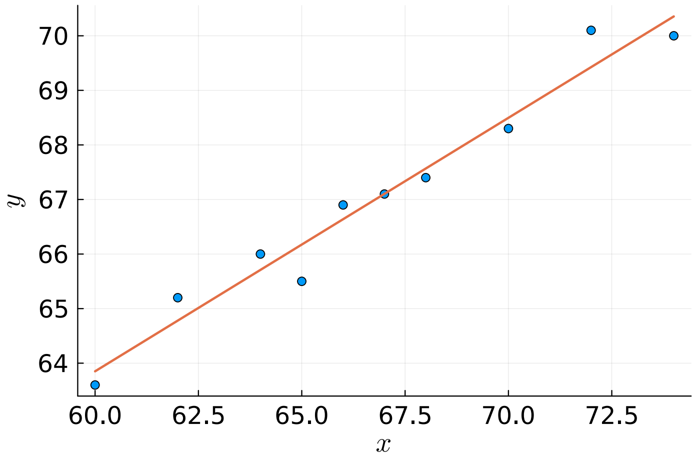
1. Konzept der Linearen Regression
2. Maximum *Likelihood* Lineare Regression
   - Methode der Kleinsten Quadrate
3. Verteilungen der Maximum *Likelihood* Schätzer
4. Hypothesentests für Regressionsparameter
   - t-Test für Lineare Regression
5. Analyse der Residuen
6. Exkurs: Multiple Regression

Überblick

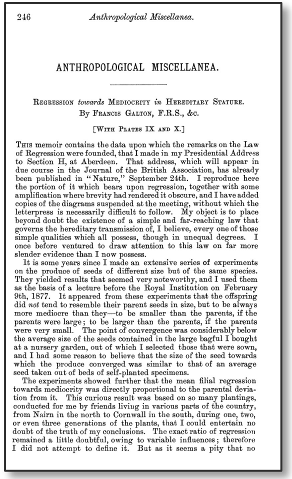
1. **Konzept der Linearen Regression**
2. Maximum *Likelihood* Lineare Regression
   - Methode der Kleinsten Quadrate
3. Verteilungen der Maximum *Likelihood* Schätzer
4. Hypothesentests für Regressionsparameter
   - t-Test für Lineare Regression
5. Analyse der Residuen
6. Exkurs: Multiple Regression

Motivation: 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?
- **Ansatz**:
  - Wir plotten die 10 Datenpunkte (x = Größe des Vaters in Inch; y = Größe des Sohnes in Inch).
  - Wir versuchen eine Gerade durch die Datenpunkte zu legen.
  - Wir testen, ob die Steigung der Gerade kleiner als 1 ist.



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

Konzepte der Linearen Regression

- **Abhängige (*response*) Variable**: Y_i
  - **Beispiel (Vererbung von Größen)**. $Y_i =$ Größe des i-ten Kindes
- **Unabhängige (*input*) Variable**: x_i
  - Keine Zufälligkeit, da diese Variablen im Experiment ausgewählt/gesetzt werden
  - **Beispiel (Vererbung von Größen)**. $x_i =$ Größe des i-ten Elternteils
- **Model**: Zusammenhang zwischen abhängigen und unabhängigen Variablen
  - Die abhängige Variable ist die Zufallsvariable.
  - Lineares Model mit einem Fehler $Z_i \sim P$ der einer bekannten Verteilung P folgt

Steigung (*slope*) $$Y_i = \boxed{\beta_1} \cdot x_i + \boxed{\beta_0} + Z_i$$ Schnitt mit der y-Achse (*intercept*)

- **Bemerkungen (Lineare Regression)**
  - Wir nehmen an, dass $E[Z_i] = 0$ (ansonsten könnten wir $Z_i{}' = Z_i - E[Z_i]$ benutzen und $E[Z_i]$ zu β_0 addieren).
  - Dieses Modell wird **lineare Regression** genannt, weil die Form der Funktion $E[Y_i] = \beta_1 \cdot x_i + \beta_0$ linear ist.

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

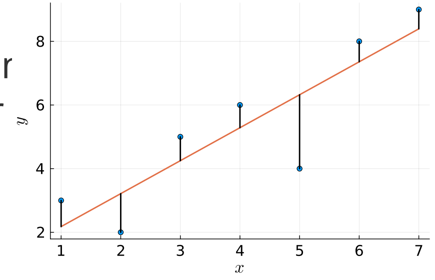
Lineare Regression mit normalverteiltem Fehler

- **Annahme**: Die Fehler sind unabhängig normalverteilt
$$Z_i \sim \mathcal{N}(0, \sigma^2)$$
- **Wiederholung (*Likelihood*)**. Ist $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ ein parametrisches Modell und sei $y \in \Omega$ eine Stichprobe, so heißt die Funktion $\mathcal{L}: \Theta \times \Omega \to [0, +\infty)$ mit $\mathcal{L}(\theta, y) = p_\theta(y)$ die zugehörige *Likelihood*, wobei p_θ die (Zähl)dichte von P_θ ist.
- **Frage**: Was sind die Parameter der Regression und wie ist die *Likelihood* definiert?
  1. Die Parameter sind
$$\theta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \sigma^2 \end{bmatrix}$$
     Schnitt mit der y-Achse (*intercept*)
     Steigung (*slope*)
     Fehlervarianz
  2. Da die Fehler unabhängig verteilt sind, gilt, dass
$$\log \mathcal{L}\big((\beta_0, \beta_1, \sigma^2), \{y_1, \dots, y_n\}\big) = \sum_{i=1}^{n} \log\big(\mathcal{N}(y_i; \beta_1 \cdot x_i + \beta_0, \sigma^2)\big)$$
$$= \sum_{i=1}^{n} -\frac{1}{2} \cdot \left[\log(2\pi\sigma^2) + \frac{\big(y_i - (\beta_1 \cdot x_i + \beta_0)\big)^2}{\sigma^2}\right]$$

Maximum *Likelihood* Lineare Regression

- **Satz (Maximum *Likelihood* Lineare Regression).** Gegeben ein lineares Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit $Z \sim \mathcal{N}(0, \sigma^2)$ und ein Datensatz von n Beobachtungen (x_i, y_i). Dann entspricht der Maximum *Likelihood* Schätzer $\beta_{0,\mathrm{MLE}}$ und $\beta_{1,\mathrm{MLE}}$ von β_0 und β_1 dem Minimierer der kleinsten Quadrate

$$\left(\beta_{0,\mathrm{MLE}}, \beta_{1,\mathrm{MLE}}\right) = \arg \min_{\beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} \left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)^2$$



- **Beweis**: Wenn wir alle Ausdrücke, die nicht von β_0 und β_1 abhängen, ignorieren, ergibt sich

$$\log \mathcal{L}\left((\beta_0, \beta_1, \sigma^2), \{y_1, \ldots, y_n\}\right) = \sum_{i=1}^{n} -\frac{1}{2} \cdot \left[\log(2\pi\sigma^2) + \frac{\left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)^2}{\sigma^2}\right]$$

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

Da der Vorfaktor strikt negativ ist, ist der Minimierer von die Maximum *Likelihood* Lösung.

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

Methode der Kleinsten Quadrate

- **Satz (Methode der Kleinsten Quadrate).** Gegeben ein Datensatz $(\boldsymbol{x}, \boldsymbol{y})$ von n Beobachtungen (x_i, y_i), sind die Minimierer $\widehat{\beta_0}$ und $\widehat{\beta_1}$ der Summe der quadratischen Abstände $\sum_{i=1}^{n}\left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)^2$ explizit darstellbar durch

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{\boldsymbol{x}} \cdot \overline{\boldsymbol{y}}}{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{\boldsymbol{x}})^2}$$

$$\widehat{\beta_0} = \overline{\boldsymbol{y}} - \widehat{\beta_1} \cdot \overline{\boldsymbol{x}}$$

- **Bemerkung (Methode der Kleinsten Quadrate).** In `R` gibt es die Funktion **`lm`** um diese Berechnung direkt auszuführen

```
x = c(1, 2, 3, 4, 5, 6, 7)
y = c(3, 2, 5, 6, 4, 8, 9)
fit = lm(y ~ x)
fit

Coefficients:
(Intercept)            x
      1.143        1.036
```

Beweis: Methode der Kleinsten Quadrate

- **Beweis**: Wir betrachten die erste Ableitung von $\sum_{i=1}^{n}\left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)^2$ nach β_0 und β_1

$$\frac{d \sum_{i=1}^{n}\left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)^2}{d\beta_0} = -2\sum_{i=1}^{n}\left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)$$

$$\frac{d \sum_{i=1}^{n}\left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)^2}{d\beta_1} = -2\sum_{i=1}^{n} x_i \cdot \left(y_i - (\beta_1 \cdot x_i + \beta_0)\right)$$

Wenn wir die erste Ableitung nach β_0 Null setzen, ergibt sich

$$0 = \sum_{i=1}^{n} y_i - \widehat{\beta_1} \cdot \sum_{i=1}^{n} x_i - n \cdot \widehat{\beta_0} \qquad \Leftrightarrow \boxed{\widehat{\beta_0} = \boldsymbol{\bar{y}} - \widehat{\beta_1} \cdot \boldsymbol{\bar{x}}}$$

Wenn wir die erste Ableitung nach β_1 zu Null setzen, ergibt sich

$$0 = \sum_{i=1}^{n} x_i \cdot y_i - \widehat{\beta_1} \cdot \sum_{i=1}^{n} x_i^2 - \widehat{\beta_0} \cdot n \cdot \boldsymbol{\bar{x}} \quad \Leftrightarrow \widehat{\beta_1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n(\boldsymbol{\bar{x}})^2\right) + n \cdot \boldsymbol{\bar{y}} \cdot \boldsymbol{\bar{x}} = \sum_{i=1}^{n} x_i \cdot y_i$$

$$\Leftrightarrow \widehat{\beta_1} = \left(\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \boldsymbol{\bar{y}} \cdot \boldsymbol{\bar{x}}\right) \Big/ \left(\sum_{i=1}^{n} x_i^2 - n(\boldsymbol{\bar{x}})^2\right)$$

Adrien-Marie Legendre
(1752 – 1833)

Carl Friedrich Gauss
(1777 - 1855)

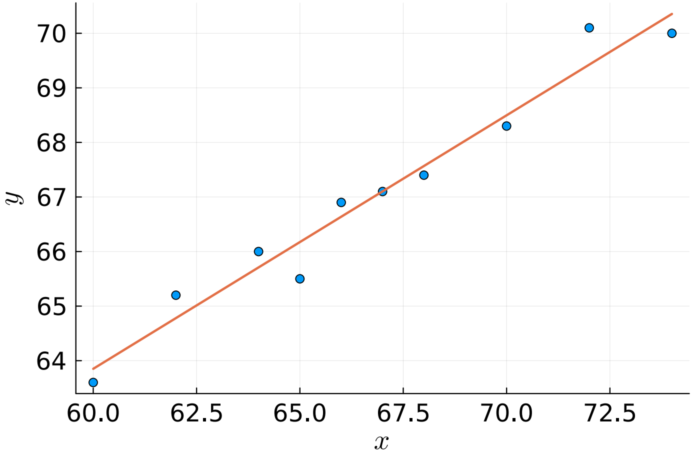
Motivation: Lineare Regression

- **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**:

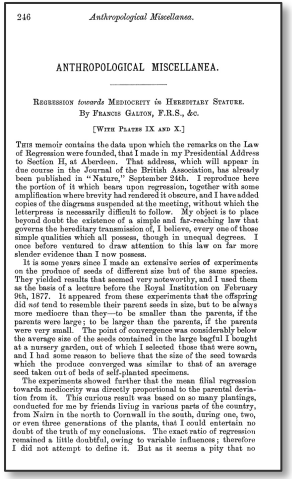
```
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    35.9768       0.4646
```

Initiale Bestätigung für $\widehat{\beta_1} < 1$.
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

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

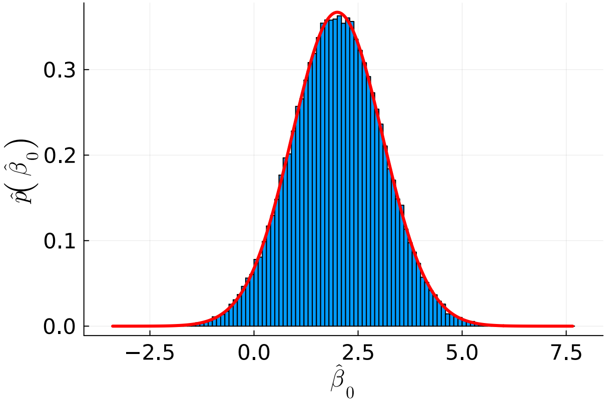
Motivation: Verteilung von Regressionsparametern

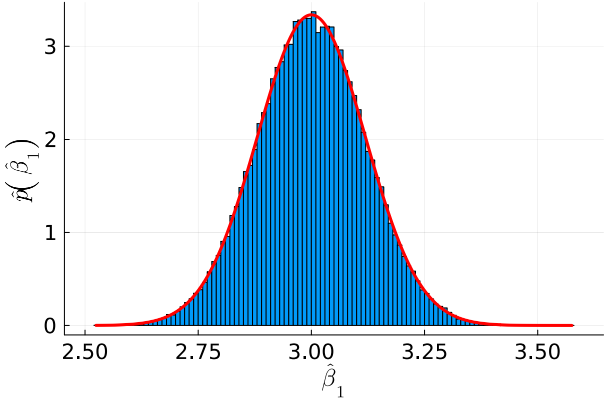
- **Beispiel (Verteilung von Regressionsparametern)**. Wir simulieren die Verteilung von $\widehat{\beta_0}$ und $\widehat{\beta_1}$ indem wir 100,000-mal $n = 15$ Datenpunkte $x_i = i$ für $\beta_0 = 2$, $\beta_1 = 3$ und $Z_i \sim \mathcal{N}(0,4)$ simulieren. Dann ergibt sich

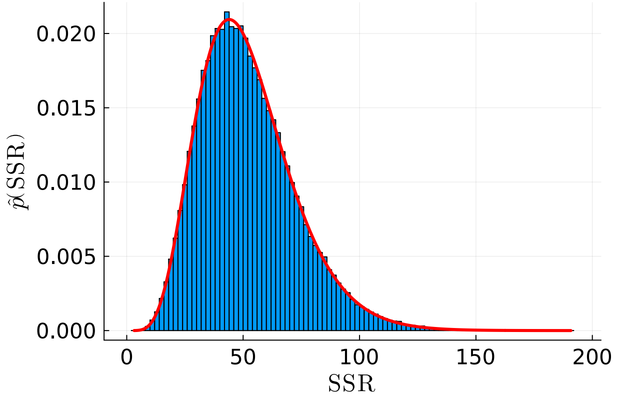
$$\widehat{\beta_0} = \bar{\boldsymbol{y}} - \widehat{\beta_1} \cdot \bar{\boldsymbol{x}}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{\boldsymbol{x}} \cdot \bar{\boldsymbol{y}}}{\sum_{i=1}^n x_i^2 - n \cdot (\bar{\boldsymbol{x}})^2}$$

$$\mathrm{SSR} = \sum_{i=1}^n \left(y_i - \left(\widehat{\beta_1} \cdot x_i + \widehat{\beta_0}\right)\right)^2$$







- Es erscheint so, dass $\widehat{\beta_0}$ und $\widehat{\beta_1}$ auch normalverteilt sind und die Summe der quadratischen Residuen χ^2-verteilt sind!
- Aber mit welchen Parametern?

Verteilung von $\widehat{\beta_1}$ (Erwartungswert)

- **Satz (Verteilung von $\widehat{\beta_1}$).** Im linearen Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit unabhängig normalverteiltem Fehler $Z \sim \mathcal{N}(0, \sigma^2)$ gilt für den MLE von β_1 für ein Stichprobe der Größe n

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

- **Beweis für $E\left[\widehat{\beta_1}\right]$** : Wir benutzen die Linearität des Erwartungswertes

$$E\left[\widehat{\beta_1}\right] = E_{Y_1,\dots,Y_n}\left[\frac{\sum_{i=1}^n x_i \cdot Y_i - n \cdot \overline{\boldsymbol{Y}} \cdot \overline{\boldsymbol{x}}}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}\right]$$

$= n \cdot \frac{1}{n} \sum_{i=1}^{n} Y_i$

$$= E_{Y_1,\dots,Y_n}\left[\frac{\sum_{i=1}^n (x_i - \overline{\boldsymbol{x}}) \cdot Y_i}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}\right]$$

da $E[Y_i] = E[\beta_1 \cdot x_i + \beta_0 + Z]$ und $E[Z] = 0$

$$= \frac{\sum_{i=1}^n (x_i - \overline{\boldsymbol{x}}) \cdot (\beta_1 x_i + \beta_0)}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}$$

$= 0$, da $\sum_{i=1}^n x_i = n \cdot \overline{\boldsymbol{x}}$ und $\sum_{i=1}^n \overline{\boldsymbol{x}} = n \cdot \overline{\boldsymbol{x}}$

$$= \frac{\beta_1 \cdot \sum_{i=1}^n x_i \cdot (x_i - \overline{\boldsymbol{x}}) + \beta_0 \cdot \sum_{i=1}^n (x_i - \overline{\boldsymbol{x}})}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2} = \beta_1 \cdot \frac{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}$$

Verteilung von $\widehat{\beta_1}$ (Varianz)

- **Satz (Verteilung von $\widehat{\beta_1}$).** Im linearen Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit unabhängig normalverteiltem Fehler $Z \sim \mathcal{N}(0, \sigma^2)$ gilt für den MLE von β_1

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

- **Beweis für $V\left[\widehat{\beta_1}\right]$**: Wir benutzen die Unabhängigkeit der Fehler

$$V\left[\widehat{\beta_1}\right] = V_{Y_1,\ldots,Y_n}\left[\frac{\sum_{i=1}^n (x_i - \overline{\boldsymbol{x}}) \cdot Y_i}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}\right]$$

Wegen Varianzzerlegungssatz

wegen Unabhängigkeit der Fehler

$$= \frac{\sum_{i=1}^n (x_i - \overline{\boldsymbol{x}})^2 \cdot V[Y_i]}{\left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)^2}$$

da $V[Y_i] = V[\beta_1 \cdot x_i + \beta_0 + Z] = V[Z]$

$$= \frac{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}{\left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)^2} \cdot \sigma^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}$$

Verteilung von $\widehat{\beta_0}$ (Erwartungswert)

- **Satz (Verteilung von $\widehat{\beta_0}$).** Im linearen Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit unabhängig normalverteiltem Fehler $Z \sim \mathcal{N}(0, \sigma^2)$ gilt für den MLE von β_0 für ein Stichprobe der Größe n

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

- **Beweis für $E[\widehat{\beta_0}]$** : Wir benutzen die Linearität des Erwartungswertes und den vorherigen Satz

$$E[\widehat{\beta_0}] = E_{Y_1,\ldots,Y_n}[\overline{Y} - \widehat{\beta_1} \cdot \overline{\boldsymbol{x}}] \qquad = \beta_1$$

$$= \frac{1}{n}\sum_{i=1}^n E[Y_i] - \beta_1 \cdot \overline{\boldsymbol{x}}$$

$= \beta_1 \cdot x_i + \beta_0$, da $E[Y_i] = E[\beta_1 \cdot x_i + \beta_0 + Z]$ und $E[Z] = 0$

$$= \frac{1}{n}\sum_{i=1}^n (\beta_1 \cdot x_i + \beta_0) - \beta_1 \cdot \overline{\boldsymbol{x}} \qquad = \beta_1 \cdot \overline{\boldsymbol{x}}$$

$$= \beta_0$$

Verteilung von $\widehat{\beta_0}$

- **Satz (Verteilung von $\widehat{\beta_0}$).** Im linearen Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit unabhängig normalverteiltem Fehler $Z \sim \mathcal{N}(0, \sigma^2)$ gilt für den MLE von β_0 für ein Stichprobe der Größe n

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

- **Beweis für $V\left[\widehat{\beta_0}\right]$** : Wir benutzen die Unabhängigkeit der Fehler und den vorherigen Satz

$$V\left[\widehat{\beta_0}\right] = V_{Y_1,\ldots,Y_n}\left[\boxed{\overline{\boldsymbol{Y}}} - \widehat{\beta_1} \cdot \overline{\boldsymbol{x}}\right]$$

$$\overline{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^n Y_i$$

$$= V_{Y_1,\ldots,Y_n}\left[\frac{1}{n}\sum_{i=1}^n Y_i\right] + (\overline{\boldsymbol{x}})^2 \cdot \boxed{V_{Y_1,\ldots,Y_n}\left[\widehat{\beta_1}\right]} + 2 \cdot \boxed{\mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^n Y_i, \widehat{\beta_1} \cdot \overline{\boldsymbol{x}}\right]}$$

$$V_{Y_1,\ldots,Y_n}\left[\widehat{\beta_1}\right] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}$$

$$\mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^n Y_i, \widehat{\beta_1} \cdot \overline{\boldsymbol{x}}\right] = \frac{\overline{\boldsymbol{x}}}{n} \cdot \sum_{i=1}^n \mathrm{Cov}\left[Y_i, \widehat{\beta_1}\right]$$

$$= \frac{\overline{\boldsymbol{x}}}{n} \cdot \sum_{i=1}^n \mathrm{Cov}\left[Y_i, \frac{\sum_{j=1}^n (x_j - \overline{\boldsymbol{x}}) \cdot Y_j}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}\right]$$

$$= \frac{\overline{\boldsymbol{x}} \cdot \left(\boxed{\sum_{i=1}^n (x_i - \overline{\boldsymbol{x}})} \cdot \sigma^2\right)}{n \cdot \left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)}$$

Mathe III

$$\sum_{i=1}^n (x_i - \overline{\boldsymbol{x}}) = 0$$

wegen Unabhängigkeit der Fehler

$$= \boxed{\frac{\sigma^2}{n}} + (\overline{\boldsymbol{x}})^2 \cdot \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2}$$

$$= \frac{\sigma^2 \cdot \left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right) + \sigma^2 \cdot n(\overline{\boldsymbol{x}})^2}{n \cdot \left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)} = \frac{\sigma^2 \cdot \sum_{i=1}^n x_i^2}{n \cdot \left(\sum_{i=1}^n x_i^2 - n(\overline{\boldsymbol{x}})^2\right)}$$

Verteilung von SSR

- **Satz (Verteilung von SSR).** Im linearen Regressionsmodell $Y = \beta_1 \cdot x + \beta_0 + Z$ mit unabhängig normalverteiltem Fehler $Z \sim \mathcal{N}(0, \sigma^2)$ gilt für die Verteilung der Summe der quadratischen Fehler bei einer Stichprobe der Größe n

$$\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2(n-2)$$

- **Bemerkungen (Verteilung von Regressionsparametern)**
  - Intuitiv ist die Anzahl der Freiheitsgrade $n-2$ weil $\mathrm{SSR}, \widehat{\beta_0}$, und $\widehat{\beta_1}$ unabhängig sind und die beiden Schätzer $\widehat{\beta_0}$ und $\widehat{\beta_1}$ jeweils einen Freiheitsgrad „entfernen".
  - Die Verteilung der Schätzer erlaubt ein Konfidenzintervall für $\widehat{\beta_0}$ und $\widehat{\beta_1}$ anzugeben, wenn wir die Varianz der Fehler als bekannt annehmen.
  - Wenn wir ein zweiseitiges Konfidenzintervall annehmen, dann gilt zum Konfidenzniveau $1-\alpha$, dass

$$\beta_1 \in \left[\widehat{\beta_1} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2 - n(\overline{\boldsymbol{x}})^2}}, \widehat{\beta_1} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2 - n(\overline{\boldsymbol{x}})^2}}\right]$$

  - Der Begriff „Regression" geht auf Sir Francis Galton zurück!

Viel Spaß bis zur nächsten Vorlesung!