VO Statistik

Sitzung 4: Aus dem Zusammenhang: Deskriptive Statistik III

Dominik Duell

Universität Innsbruck

Etwas Admin

Pop Quiz

- Mehr Fragen, weil Auswahl an Büchern
- Folge dem Code, der an die Wand projeziert ist

Plan für heute

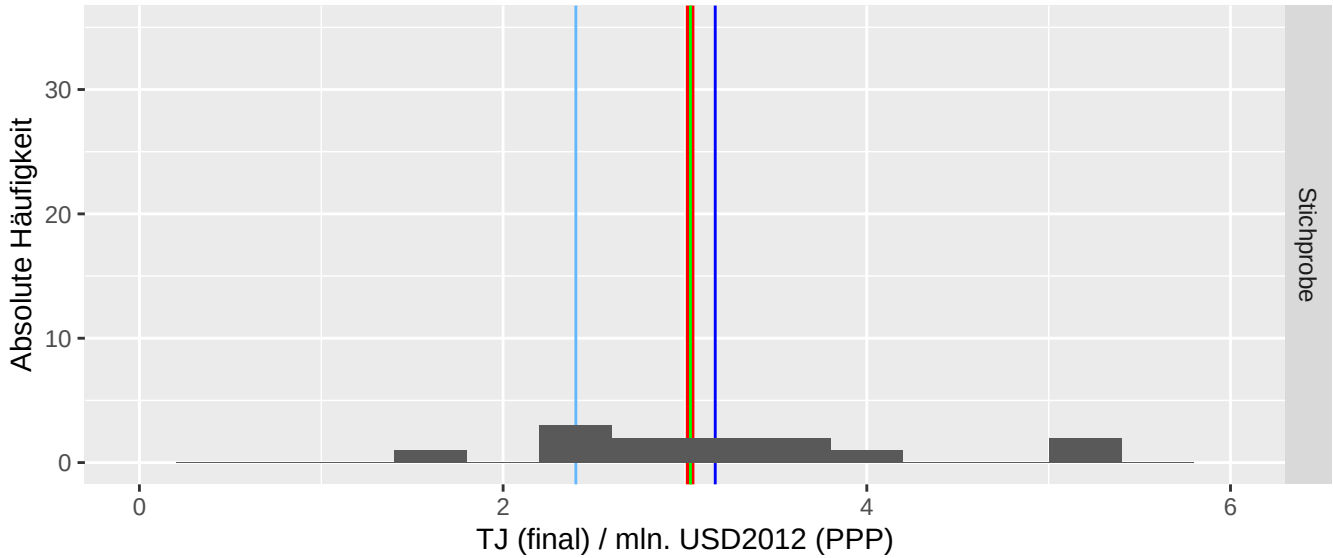
1. Statistische Modell zu Ende gedacht
2. Streuungsmaße (Varianz, Dispersion, Quantilabstand etc.)
3. Modelle statistischer Zusammenhänge

→ Bei den Fehlern sprechen wir auch von Abweichungen

Statistische Modelle

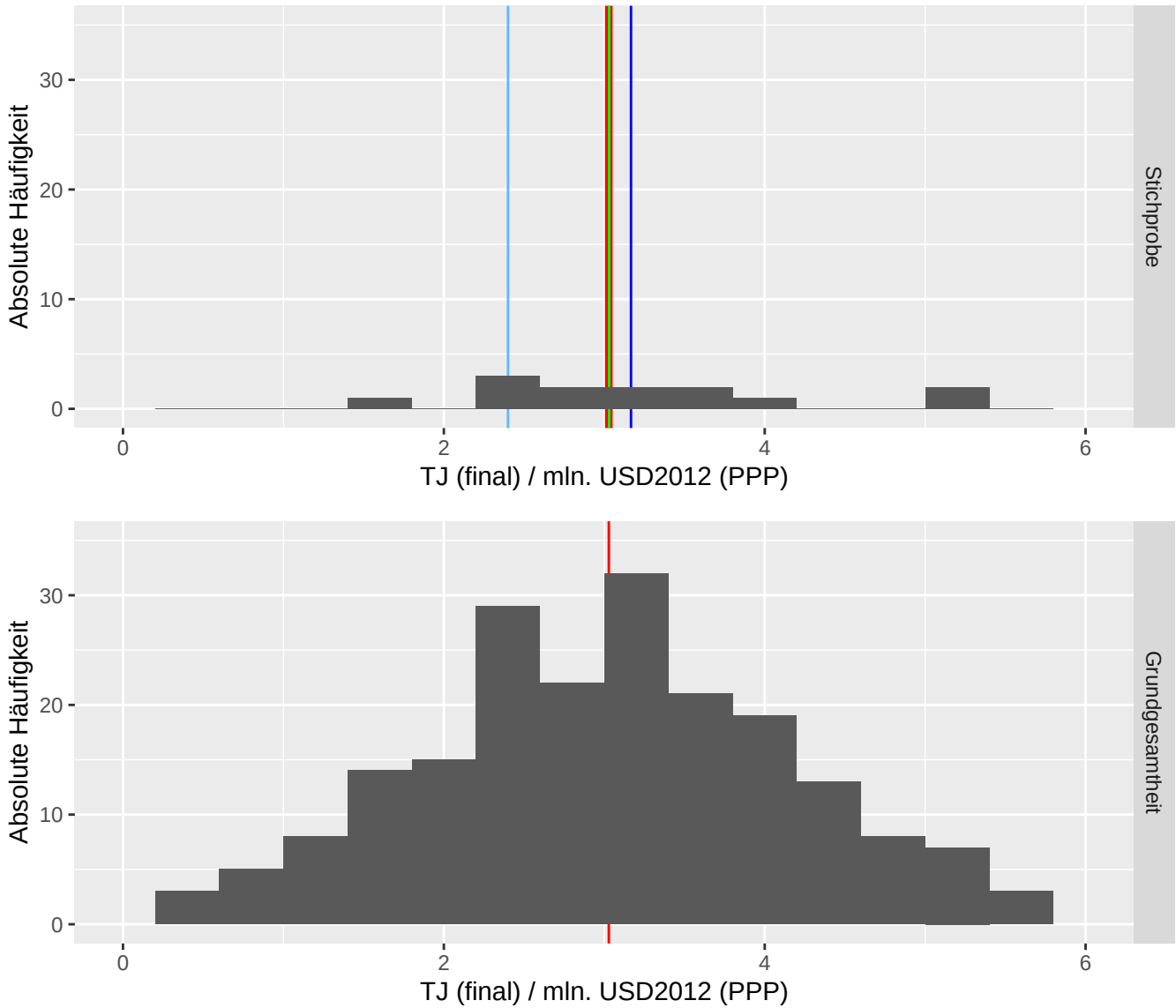
→ Bei den Fehlern sprechen wir auch von Abweichungen

Grundgesamtheit $\neq$ Stichprobe



→ Bei den Fehlern sprechen wir auch von Abweichungen

Grundgesamtheit ≠ Stichprobe



→ Bei den Fehlern sprechen wir auch von Abweichungen

Modelliere, modelliere

$$outcome_i = model + error_i$$

→ Bei den Fehlern sprechen wir auch von Abweichungen

Modelliere, modelliere

$$outcome_i = model + error_i$$

$$emissions_i = \overline{emissions} + error_i$$

→ Bei den Fehlern sprechen wir auch von Abweichungen

Modelliere, modelliere

$$outcome_i = model + error_i$$

$$emissions_i = \overline{emissions} + error_i$$

$$emissions_i = b_0 + error_i$$

→ Bei den Fehlern sprechen wir auch von Abweichungen

Modelliere, modelliere

$$outcome_i = model + error_i$$

$$emissions_i = \overline{emissions} + error_i$$

$$emissions_i = b_0 + error_i$$

$$emissions_i = b_0 + b_1 GDP_i + error_i$$

→ Bei den Fehlern sprechen wir auch von Abweichungen

Modelliere, modelliere

$$outcome_i = model + error_i$$

$$emissions_i = \overline{emissions} + error_i$$

$$emissions_i = b_0 + error_i$$

$$emissions_i = b_0 + b_1 GDP_i + error_i$$

→ Bei den Fehlern sprechen wir auch von Abweichungen

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist der Fehler

$$emissions_i = \overline{emissions} + error_i$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist der Fehler

$$emissions_i = \overline{emissions} + error_i$$

$$error_i = emissions_i - \overline{emissions}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist der Fehler

$$emissions_i = \overline{emissions} + error_i$$

$$error_i = emissions_i - \overline{emissions}$$

$$error_i = emissions_i - b_0$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist der Fehler

$$emissions_i = \overline{emissions} + error_i$$

$$error_i = emissions_i - \overline{emissions}$$

$$error_i = emissions_i - b_0$$

$$error_i = emissions_i - b_0 - b_1 GDP_i$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist der Fehler

$$emissions_i = \overline{emissions} + error_i$$

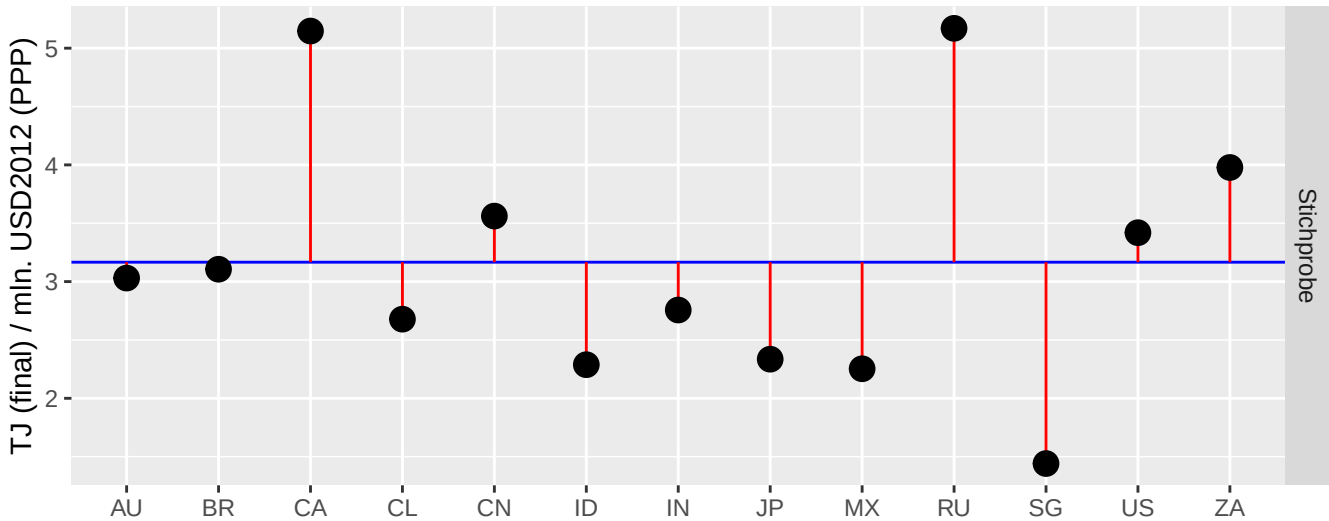
$$error_i = emissions_i - \overline{emissions}$$

$$error_i = emissions_i - b_0$$

$$error_i = emissions_i - b_0 - b_1 GDP_i$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Wo ist der Fehler



→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} error_i$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} error_i$$

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} (emissions_i - \overline{emissions})$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} error_i$$

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} (emissions_i - \overline{emissions})$$

$$\text{Summe der quadrierten Abweichungen} = \sum_{i=1}^{N} (emissions_i - \overline{emissions})^2$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} error_i$$

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} (emissions_i - \overline{emissions})$$

$$\text{Summe der quadrierten Abweichungen} = \sum_{i=1}^{N} (emissions_i - \overline{emissions})^2$$

Ob positive oder negative Abweichung darf nichts ausmachen.

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Mittlere quadrierte Abweichung} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Mittlere quadrierte Abweichung} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

Um so mehr Beobachtungen, umso geringer der Mean squared error

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Mittlere quadrierte Abweichung} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

Um so mehr Beobachtungen, umso geringer der Mean squared error

Das ist die **Varianz** von *emissions*

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Was ist ein gutes Modell

$$\text{Mittlere quadrierte Abweichung} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

Um so mehr Beobachtungen, umso geringer der Mean squared error

Das ist die **Varianz** von *emissions*

Mittlere quadrierte Abweichung = Varianz wenn das Modell der Mittelwert ist.

→ Bei den Fehlern sprechen wir auch von Abweichungen

Varianz und Standardabweichung

Varianz von *emissions*

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Varianz und Standardabweichung

Varianz von *emissions*

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

Standardabweichung (standard deviation)

$$\sigma = \sqrt{\sigma^2} = \text{sd(emissions)}$$

→ Bei den Fehlern sprechen wir auch von Abweichungen

Varianz und Standardabweichung

Varianz von *emissions*

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

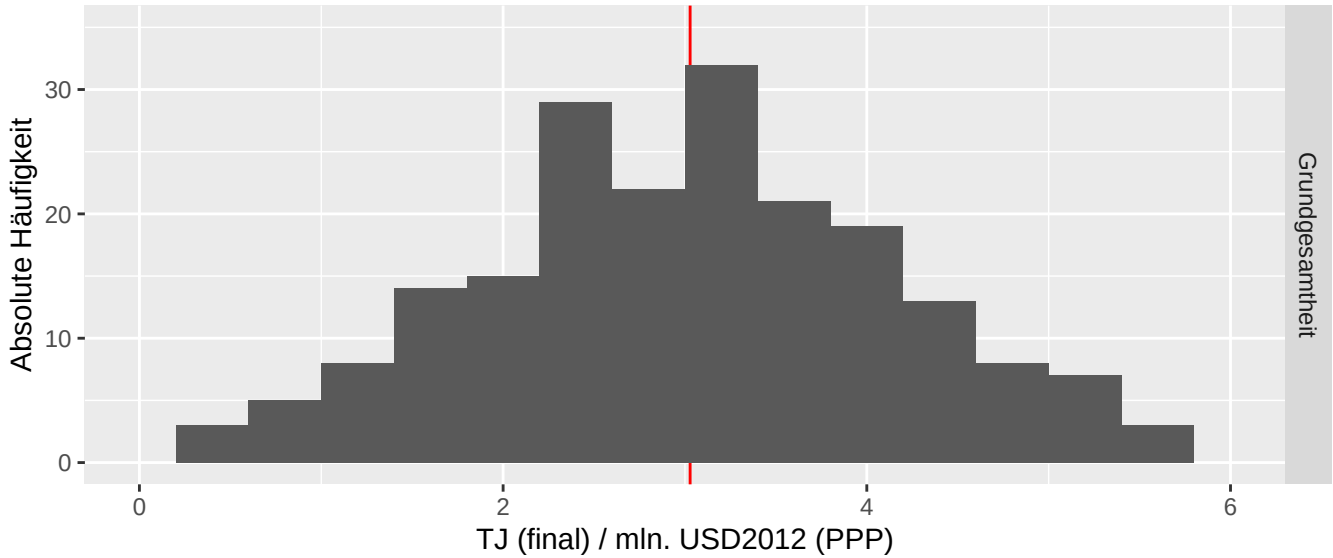
Standardabweichung (standard deviation)

$$\sigma = \sqrt{\sigma^2} = \text{sd(emissions)}$$

Mit der Standardabweichung kommen wir zur ursprünglichen Einheit der Variable zurück.

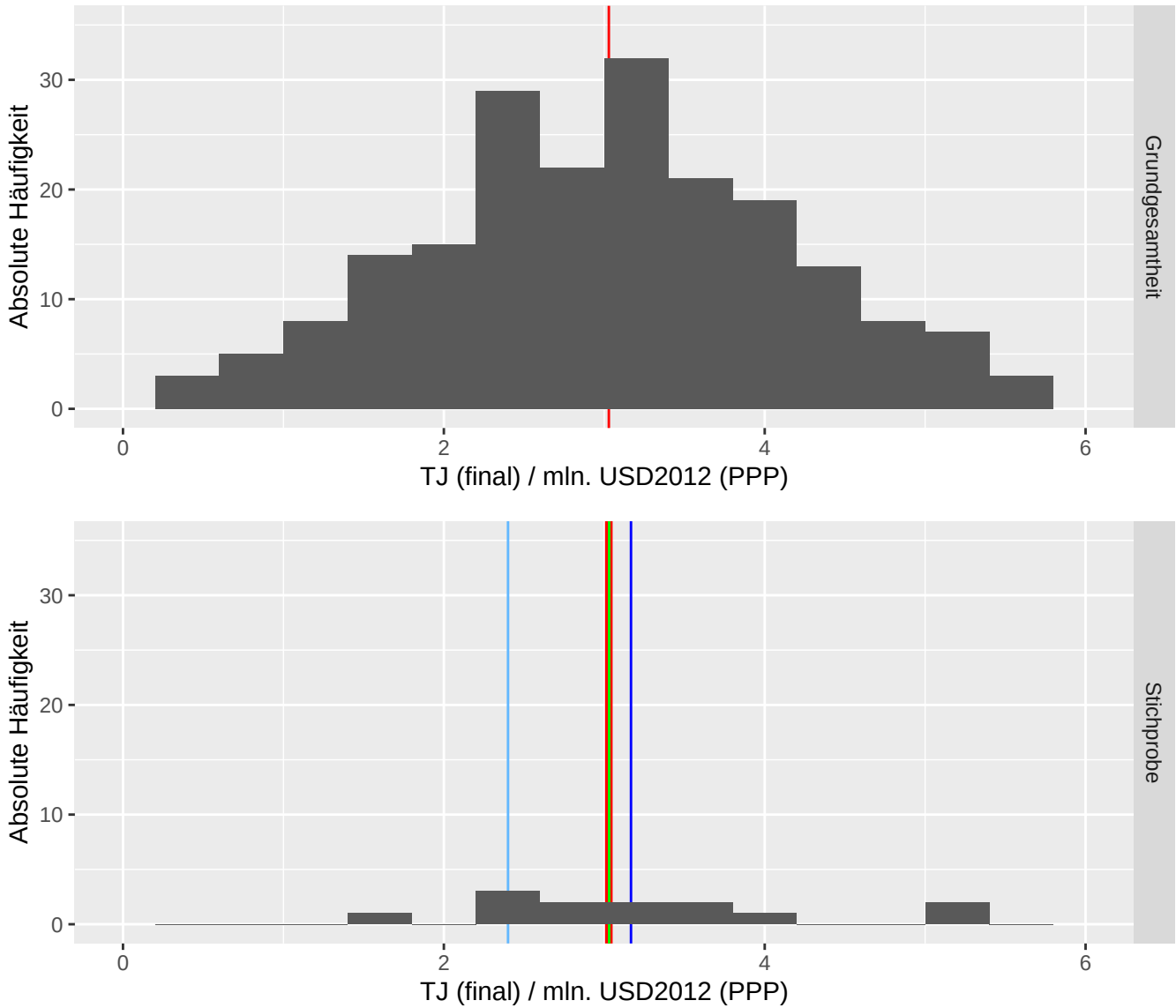
→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Grundgesamtheit ≠ Stichprobe



→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Grundgesamtheit ≠ Stichprobe



→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Statistiken für Grundgesamtheit und Stichprobe

Grundgesamtheit:

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Statistiken für Grundgesamtheit und Stichprobe

Grundgesamtheit:

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

$$\sigma = \sqrt{\sigma^2} = \text{Standardabweichung(emissions)}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Statistiken für Grundgesamtheit und Stichprobe

Grundgesamtheit:

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

$$\sigma = \sqrt{\sigma^2} = \text{Standardabweichung(emissions)}$$

$$s^2 = var(emissions)^{\text{Stichprobe}} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{\mathbf{N-1}}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Statistiken für Grundgesamtheit und Stichprobe

Grundgesamtheit:

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

$$\sigma = \sqrt{\sigma^2} = \text{Standardabweichung(emissions)}$$

$$s^2 = var(emissions)^{\text{Stichprobe}} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{\mathbf{N-1}}$$

$$s = \sqrt{s} = \text{Standardabweichung}^{\text{Stichprobe}}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Statistiken für Grundgesamtheit und Stichprobe

Grundgesamtheit:

$$\sigma^2 = var(emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

$$\sigma = \sqrt{\sigma^2} = \text{Standardabweichung(emissions)}$$

$$s^2 = var(emissions)^{\text{Stichprobe}} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{\mathbf{N-1}}$$

$$s = \sqrt{s} = \text{Standardabweichung}^{\text{Stichprobe}}$$

→ **Bei den Fehlern sprechen wir auch von Abweichungen**

Streuungsmaße

Varianz und Standardabweiung

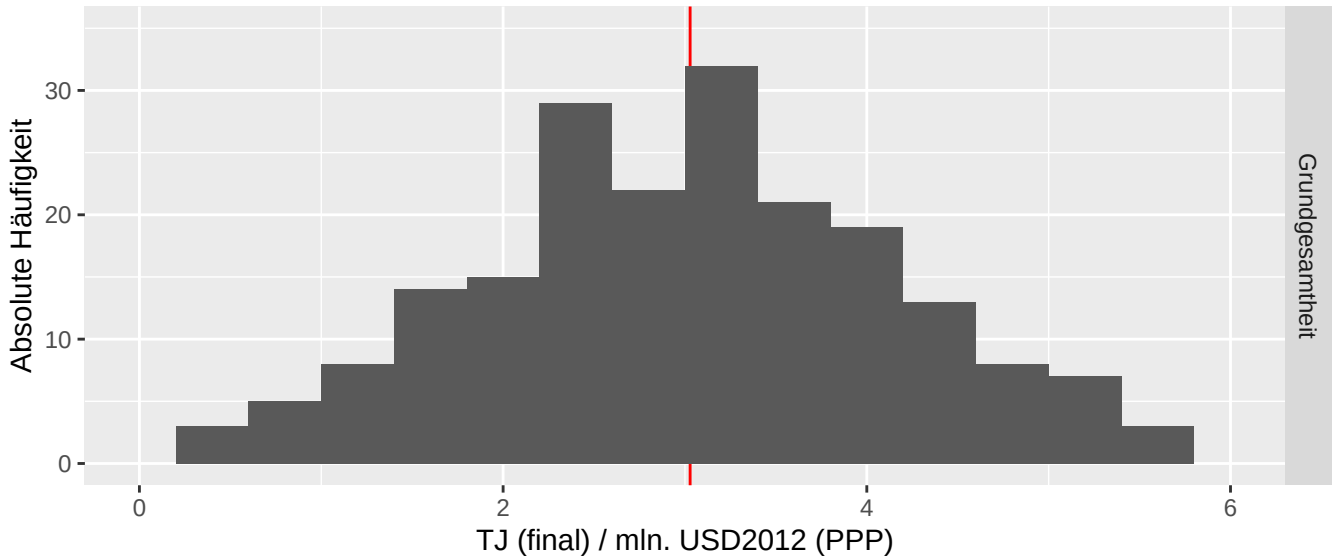
$$\sigma^2 = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

Varianz und Standardabweiung

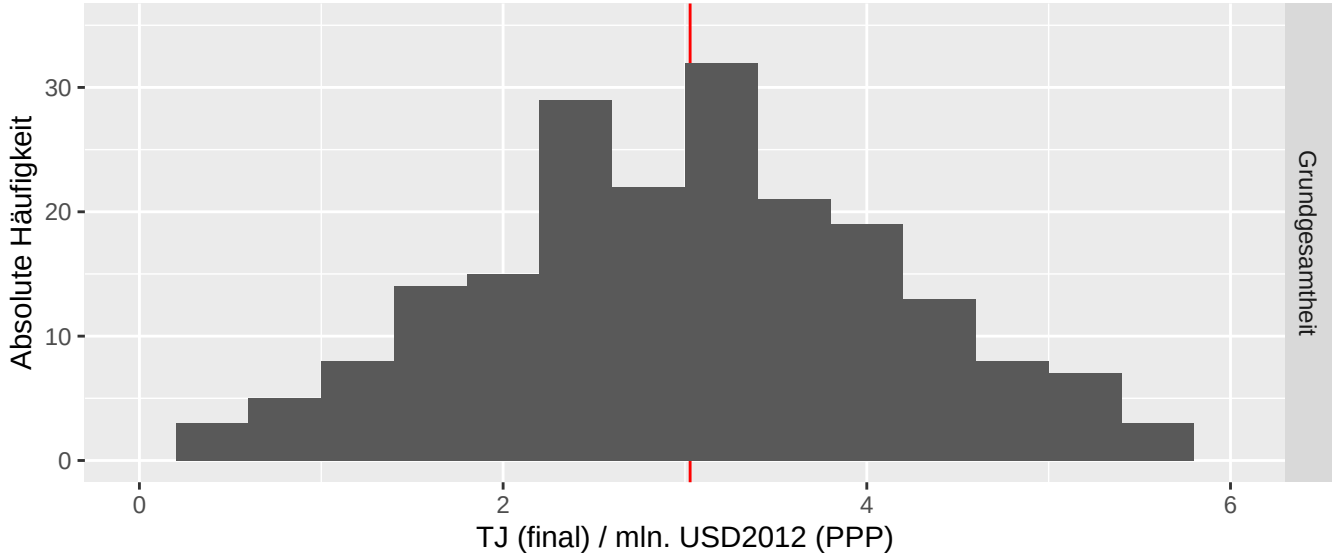
$$\sigma^2 = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

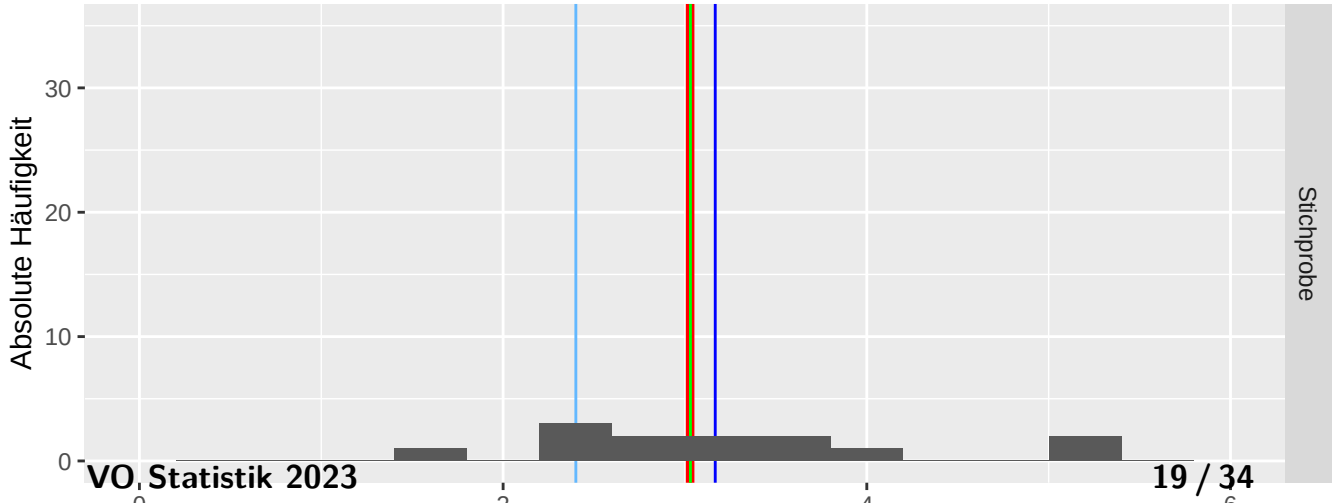
$$\sigma = \sqrt{\sigma^2}$$

Wie sehen verschieden große Standardabweichungen aus?



Wie sehen verschieden große Standardabweichungen aus?





Weiter Streuungsmaße: Spannweite

$$max(emissions_i) - min(emissions_i)$$

oder wir ordnen die Werte der Variable *emissions* der Größe nach und nennen den kleinsten Wert $emissions_1$ und den größten Wert $emissions_N$ für N Werte.

$$emissions_N - emissions_1$$

Weiter Streuungsmaße: Interquartilsabstand

- Nach Ordnung der Werte der Variable von kleinstem zu größten Wert bestimmen wir **Quartile**:
  - **0,25 Quartile**: der Wert der Variable so dass 25% der Werte kleiner und 75% der Werte größer sind:

Weiter Streuungsmaße: Interquartilsabstand

- Nach Ordnung der Werte der Variable von kleinstem zu größten Wert bestimmen wir **Quartile**:
  - **0,25 Quartile**: der Wert der Variable so dass 25% der Werte kleiner und 75% der Werte größer sind: $emissions_i^{0,25}$

Weiter Streuungsmaße: Interquartilsabstand

- Nach Ordnung der Werte der Variable von kleinstem zu größten Wert bestimmen wir **Quartile**:
  - **0,25 Quartile**: der Wert der Variable so dass 25% der Werte kleiner und 75% der Werte größer sind: $emissions_i^{0,25}$
  - **0,75 Quartile**: der Wert der Variable so dass 75% der Werte kleiner und 25% der Werte größer sind:

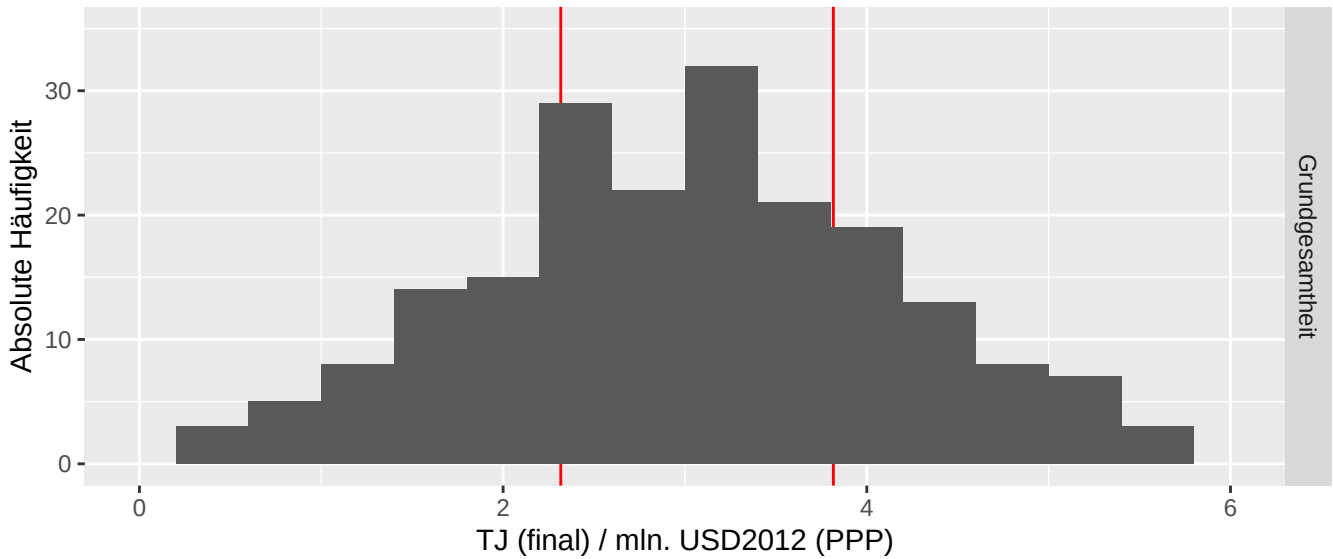
Weiter Streuungsmaße: Interquartilsabstand

- Nach Ordnung der Werte der Variable von kleinstem zu größten Wert bestimmen wir **Quartile**:
  - **0,25 Quartile**: der Wert der Variable so dass 25% der Werte kleiner und 75% der Werte größer sind: $emissions_i^{0,25}$
  - **0,75 Quartile**: der Wert der Variable so dass 75% der Werte kleiner und 25% der Werte größer sind: $emissions_i^{0,75}$

Interquartilsabstand:

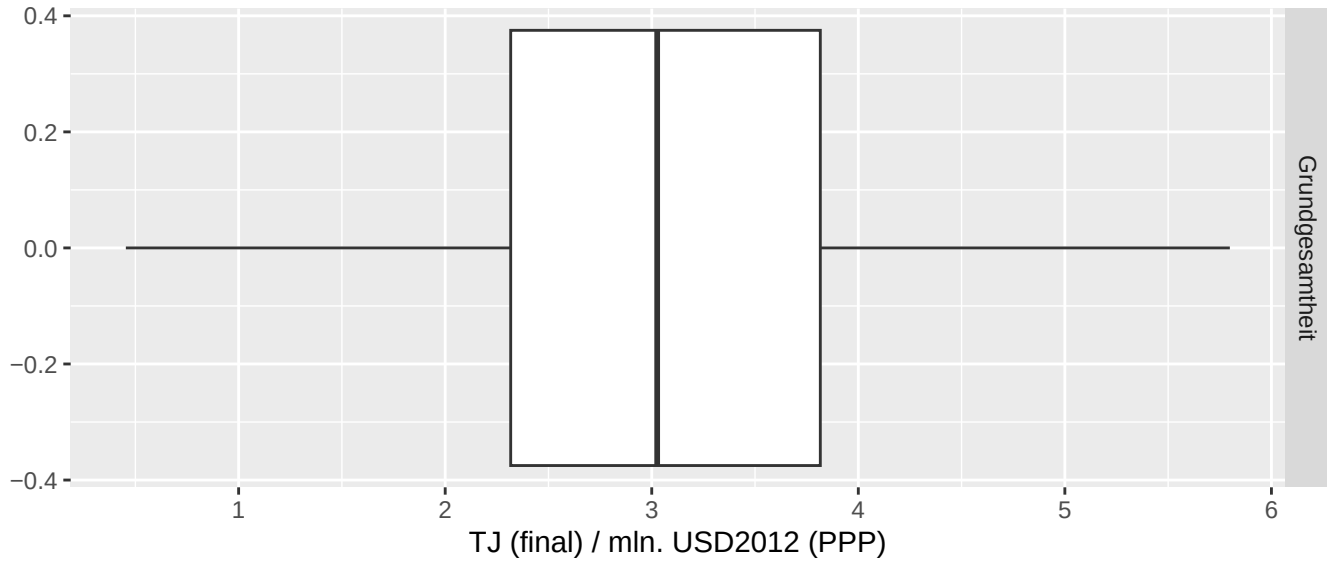
$$emissions_i^{0,75} - emissions_i^{0,25}$$

Interquartilsabstand



→ Wie der Median ist der Interquartilsabstand nicht leicht durch extreme Werte zu beeinflussen.

Interquartilsabstand grafisch? Boxplot!



Modelle statistischer Zusammenhänge

Geht das eine hoch, geht auch das andere hoch?

Kovarianz von *emissions* und *GDP*:

$$cov(emissions, GDP) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions_i})(GDP_i - \overline{GDP}_i)}{N-1}$$

Mache Kovarianz vergleichbar

Pearson's Korrelationskoeffizent

$$\rho_{emissions,GDP} = \frac{cov(emissions, GDP)}{s_{emissions} s_{GDP}}$$

ρ ist zwischen 0 und 1 für die Korrelation zwischen jeglichen Variablen.

Einfaches lineares Regressionsmodell

$$emissions_i = b_0 + b_1 GDP_i + u_i$$

- b_0 sagt uns wo unsere Schätzung liegt im Verhältnis zu *emissions*
- b_1 gibt uns die Form, wie *emissions* und *GDP* zu einander stehen.
- b_0 und b_1 bezeichnen wir als Regressionskoeffizient.

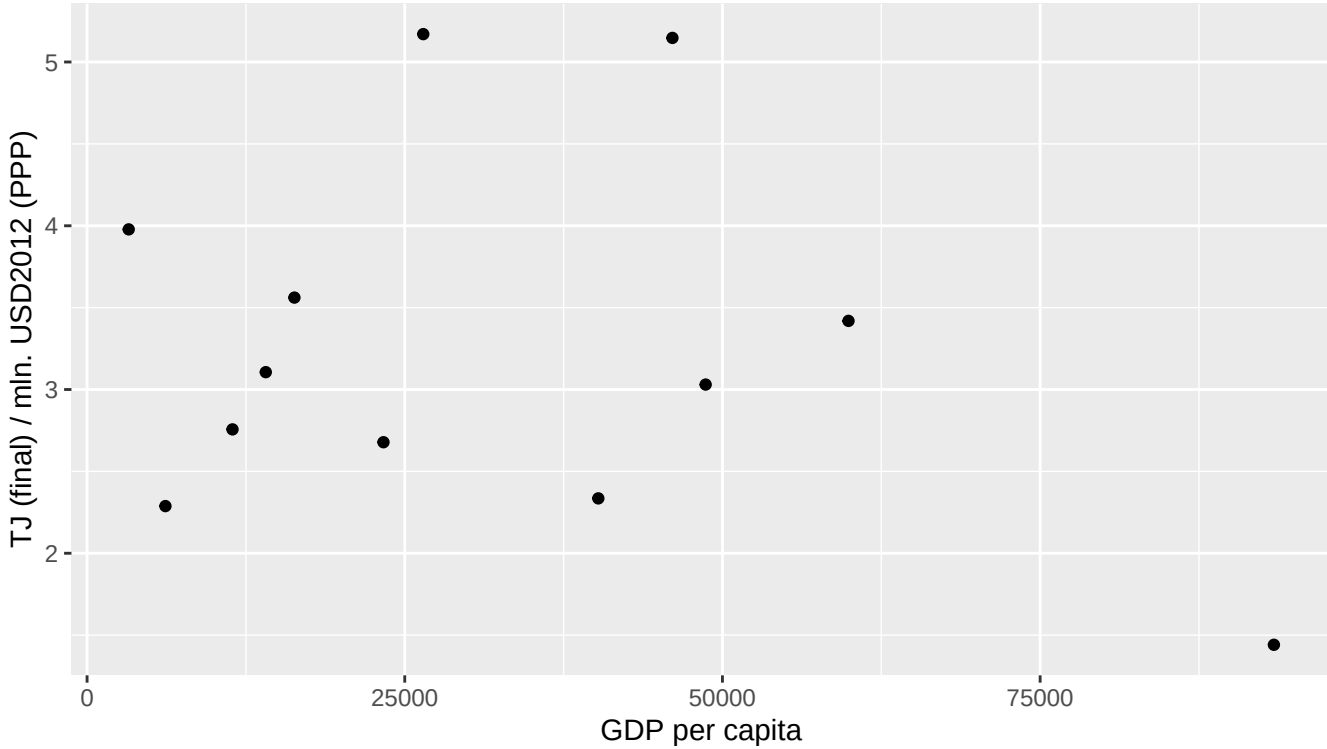
Einfaches lineares Regressionsmodell

Definition: Das lineare Modell

$$y_i = b_0 + b_1 x_i + u_i$$

heißt einfaches lineares Regressionsmodell. Dabei bezeichnet die Variable X die unabhängige Variable, auch Regressor oder erklärende Variable genannt. Die abhängige Variable Y heißt Regressand, erklärte oder zu erklärende Variable. Die Fehler u_i beschreiben die möglichen Abweichungen der Gerade von den Beobachtungen, da bis auf wenige Ausnahmen die Beobachtungen nicht auf der Geraden liegen werden.
Quelle: Sibbertsen/Lehne, S.137

Emissionen und Wirtschaftsleistung

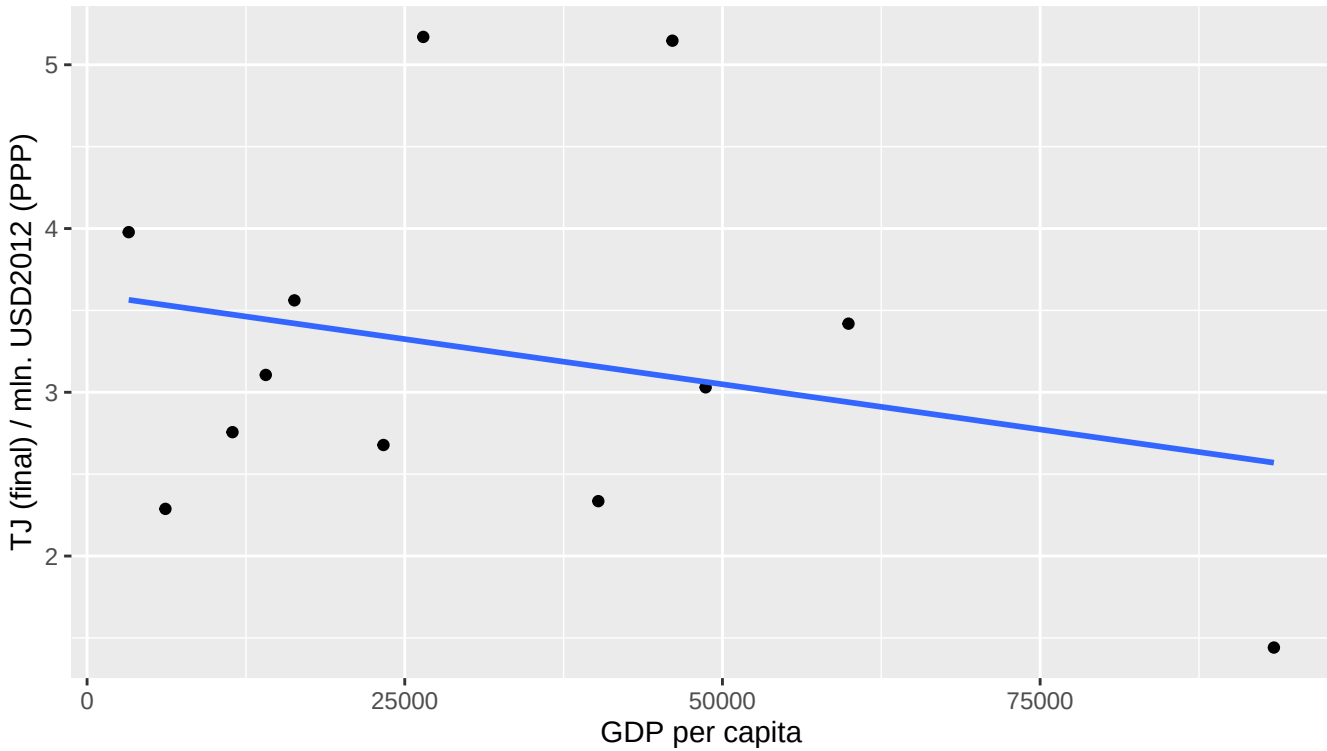


$\rightarrow$ Zusammenhang?

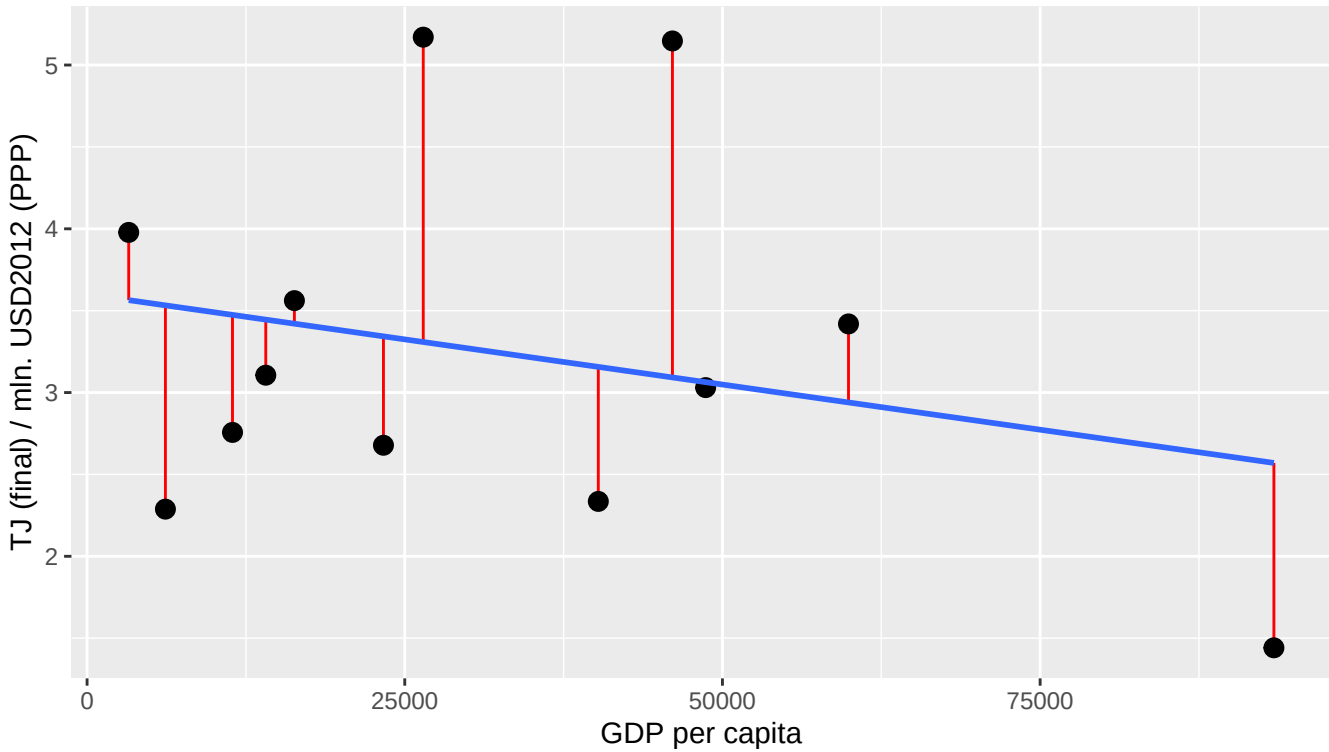
Kovarianz: -7701.71

Korrelation: -0.26

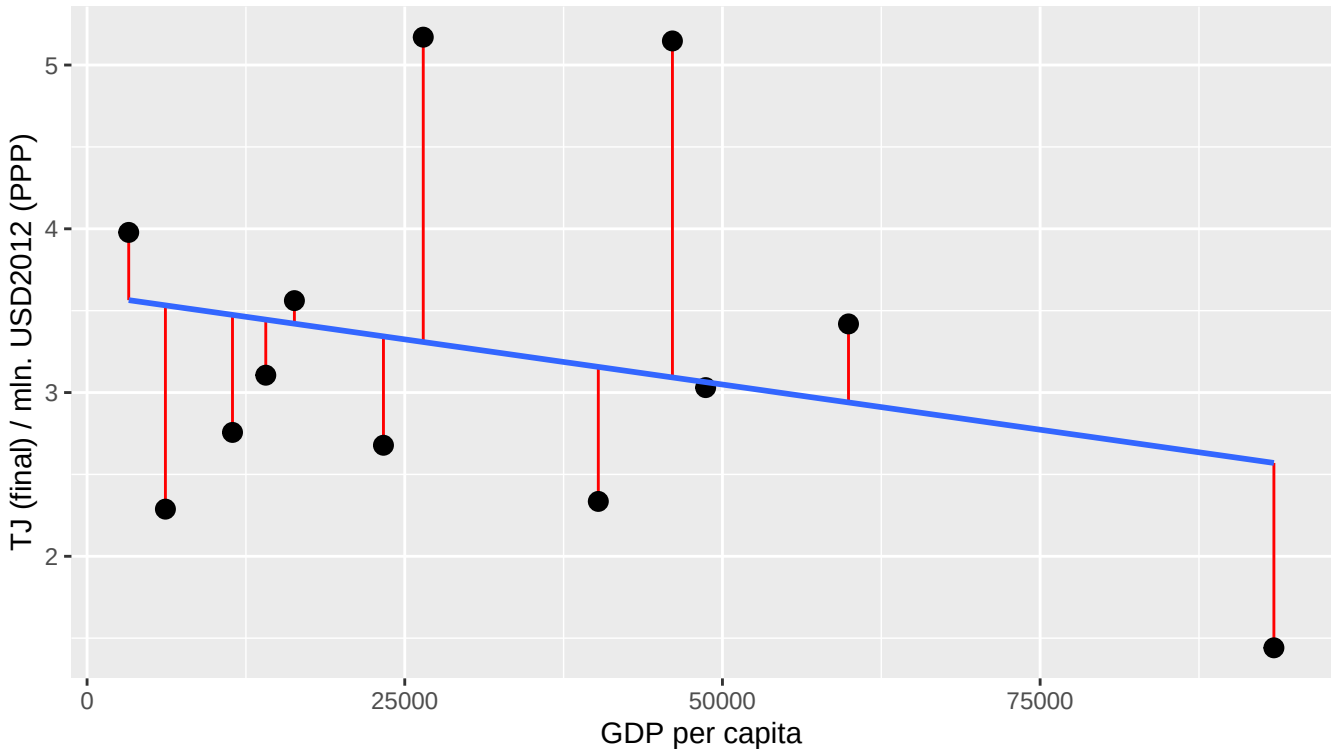
Emissionen und Wirtschaftsleistung



Was ist unsere beste Schätzung?

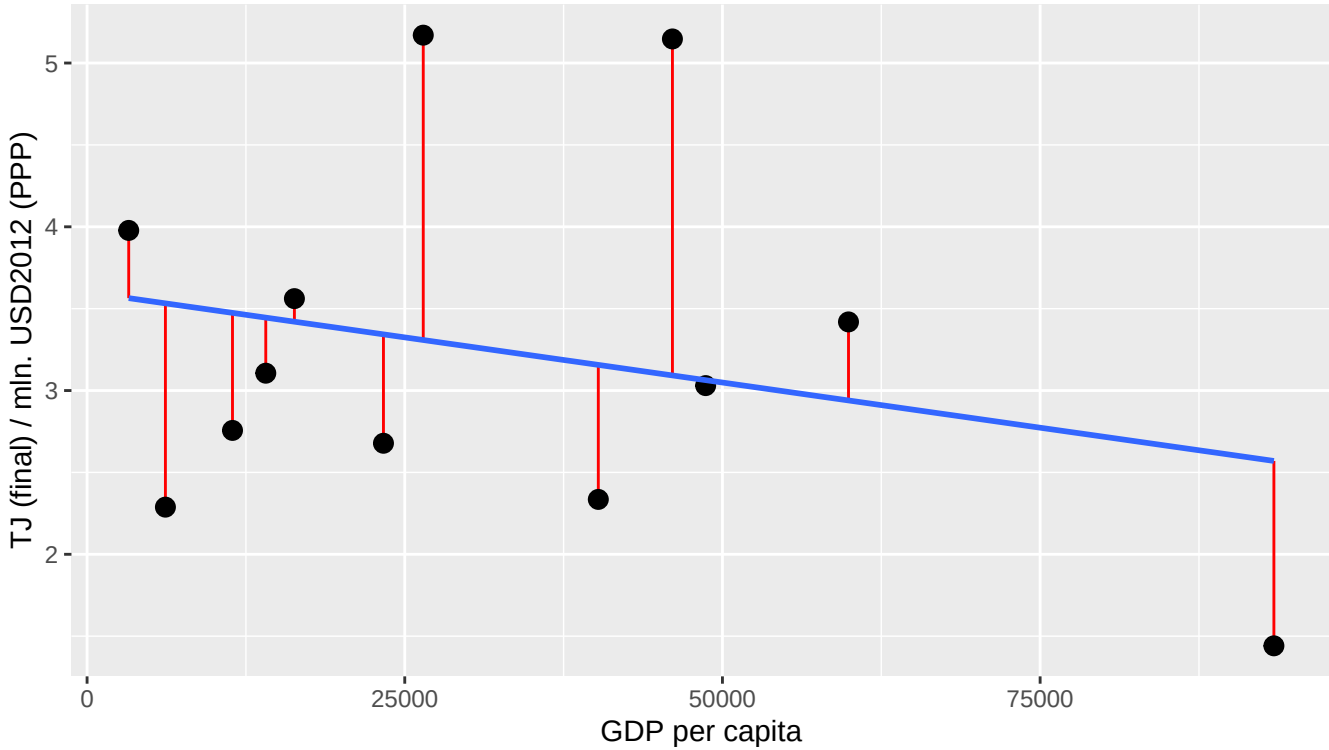


Was ist unsere beste Schätzung?



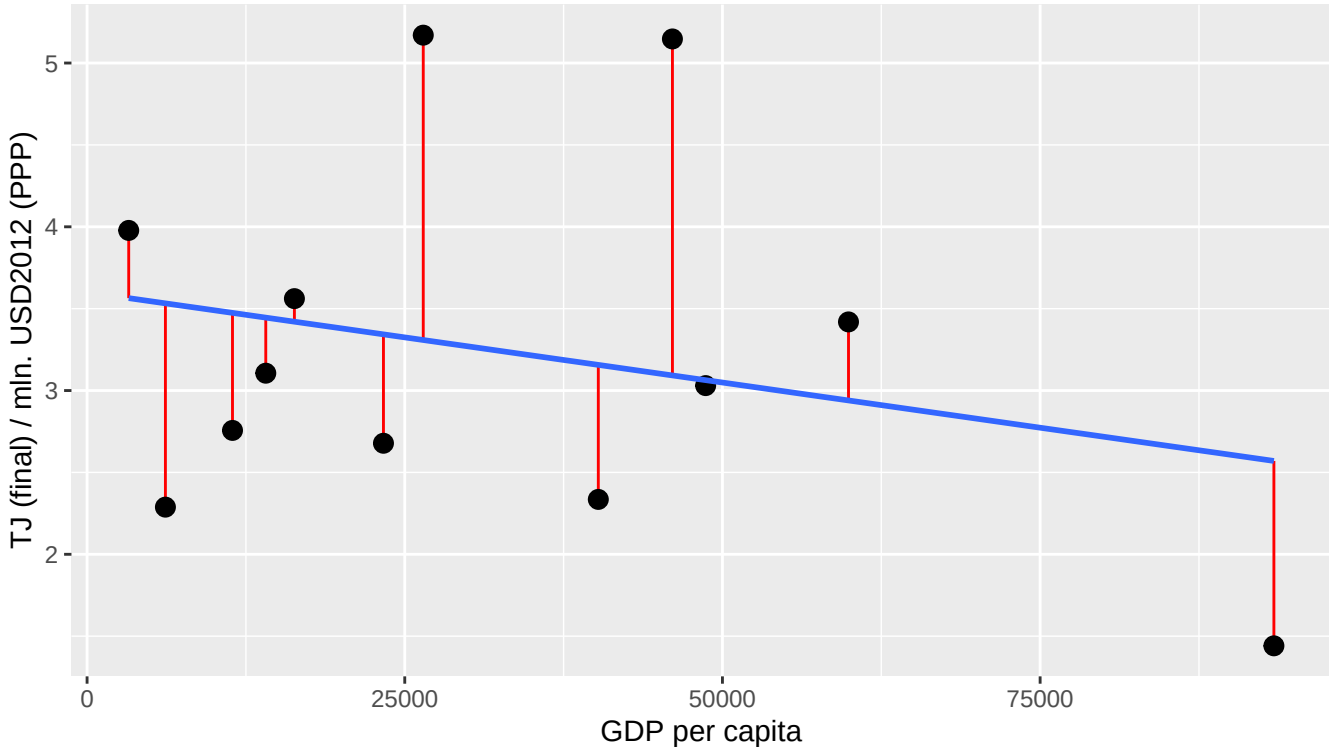
Die Linie, welche die Summe der quadrierten Fehler minimiert.

Was ist unsere beste Schätzung?



Die Linie, welche die Summe der quadrierten Fehler minimiert.
Methode der kleinsten Quadrate

Was ist unsere beste Schätzung?



Die Linie, welche die Summe der quadrierten Fehler minimiert.
Methode der kleinsten Quadrate

Für unser lineares Modell

$$emissions_i = b_0 + b_1 GDP_i + \epsilon_i$$

ist die Steigung dieser Linie (unsere beste Schätzung von b_1) immer gegeben als:

$$\hat{b}_1 = \frac{Cov(emissions, GDP)}{Var(emissions)}$$

Und der beste Schätzer for den Achsenabschnitt b_0 ist

$$\hat{b}_0 = \hat{y} - \hat{b}_1 \overline{emissions}$$

Was solltet ihr von heute mitnehmen?

1. Verstehen, dass wir ein Modell brauchen, um mit den Daten aus unserer Stichprobe Statistiken der Grundgesamtheit zu schätzen
2. Wissen was grundlegene Modelle sind, welche ganz gute Annäherungen der Grundgesamtheit sind (solange wir eine Zufallsstichprobe gezogen und genug Beobachtungen zu haben)
3. Kenntniss grundlegender Streuungsmaße
4. Wissen, wie man einen (linearen) Zusammenhang gut schätzen kann

Was gibt's nächste Woche

1. Wahrscheinlichkeitsrechnung
2. Unsicherheit in statistischen Modellen
3. Verteilungen der Stichprobenstatistiken