

Start:Bayes!

Einfache lineare Modelle

Sebastian Sauer



Einstieg



Lernziele

Nach diesem Kapitel können Sie ...

- das Zusammenspiel von Apriori-Verteilungen und linearem Modell zur Berechnung der Likelihood erläutern
- die Post-Verteilung für einfache lineare Modelle in R berechnen
- zentrale Informationen (Lage-, Streuungsmaße, Schätzintervalle) aus der Post-Verteilung eines Regressionsmodells herauslesen

Orientiert an McElreath ([2020](#)), Kap. 4.4.



Überblick

- Modell: Körpergröße vorhergesagt durch Gewicht (!Kung-Datensatz)
- Neu ggü. Frequentistischer Regression (**lm**): Die drei Parameter β_0 , β_1 , σ haben jetzt eine **Post-Verteilung**
- Frequentistisch: nur Punktschätzer pro Parameter
- Bayes: eine ganze Verteilung pro Parameter — informationsreicher

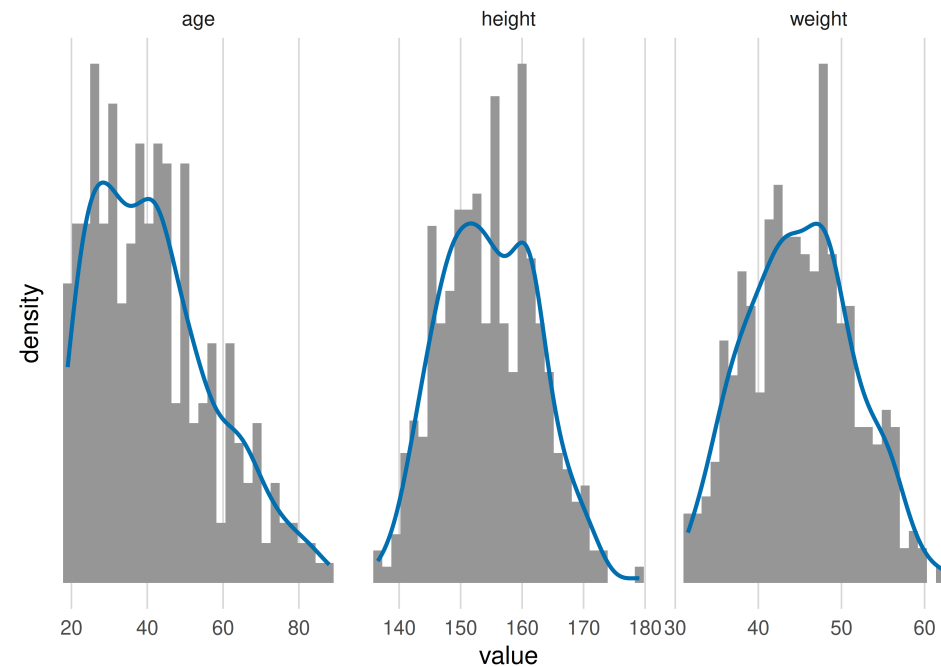


Den Datensatz verstehen



Der Datensatz: !Kung-Leute

- Datei `Howell1a.csv` ([McElreath, 2020](#))
- Nur Erwachsene (Alter > 18)
- Variablen u. a.: `height` (Größe), `weight` (Gewicht), `age`, `male`

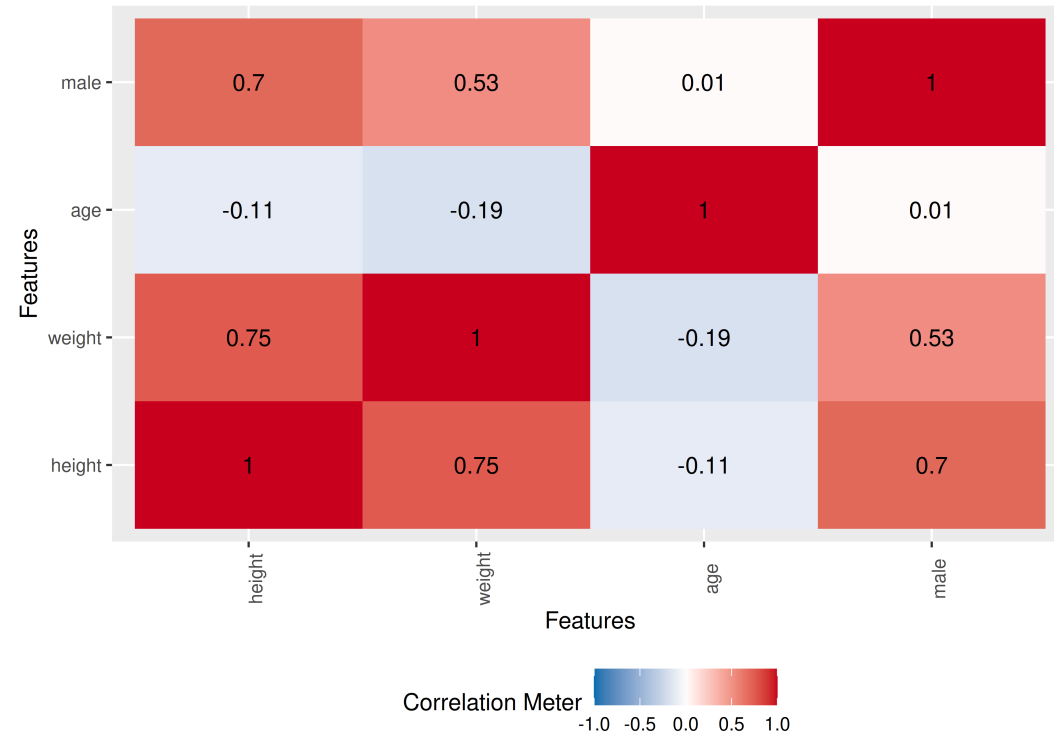


Explorative Datenanalyse (EDA)

- Nicht zwingend nötig für die Bayes-Regression, aber hilfreich zum Datenverständnis
- Keine fehlenden Werte im Datensatz
- `age` rechtsschief; `height` und `weight` eher symmetrisch/normalverteilt
- `height` bimodal — vermutlich Mischung der Geschlechter
- Etwas mehr Männer als Frauen im Datensatz



Zusammenhänge in den Daten



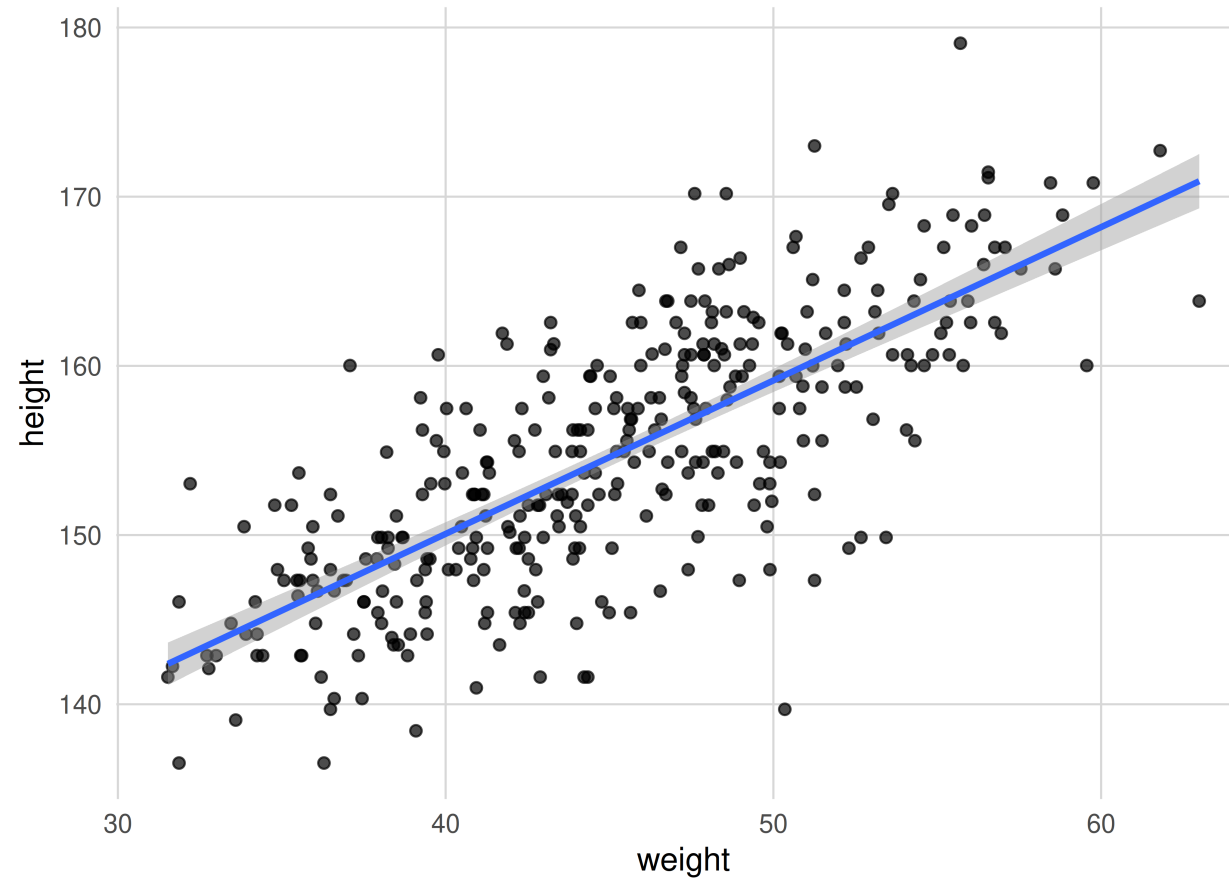
Starke Zusammenhänge zwischen Größe und Gewicht sowie zwischen Geschlecht und Größe/Gewicht.

Gewicht und Größe: linearer Zusammenhang

- Regression prüft, inwieweit zwei Variablen Y und X linear zusammenhängen
- Je stärker der Zusammenhang, desto besser lässt sich X zur Vorhersage von Y nutzen
- Zusammenhang $\neq$ Kausalität
- *Linear*: Zuwachs in Y relativ zu X ist konstant — steigt X um 1 Einheit, steigt Y um b Einheiten



Gewicht (X) und Größe (Y)



Deutlicher linearer Zusammenhang: je schwerer, desto größer.

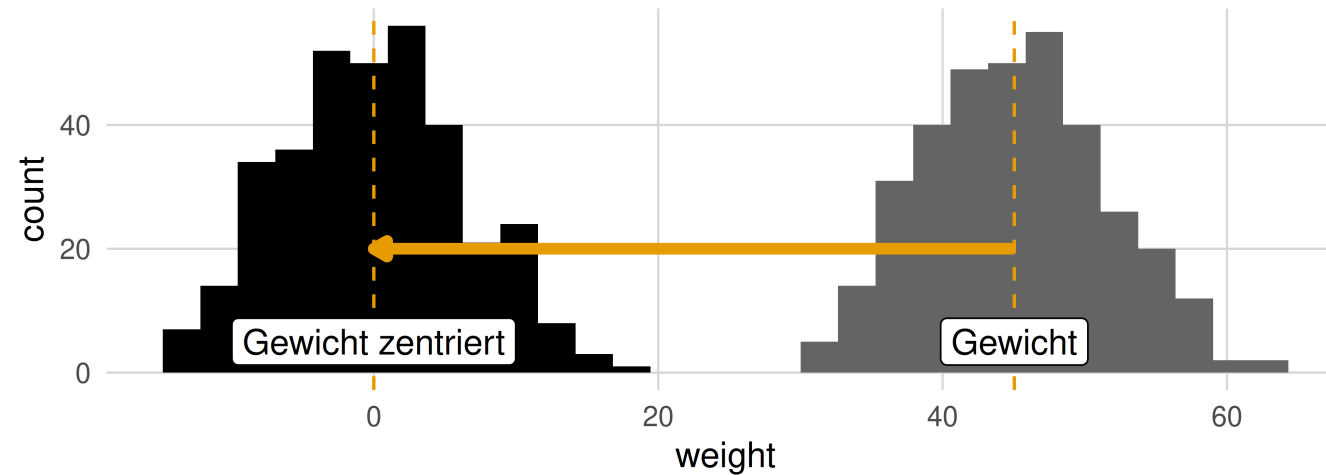
Prädiktor zentrieren

- Zentrieren: von jedem Wert den Mittelwert abziehen —
 $x_i - \bar{x}$
- Ergebnis: `weight_c` (c wie *centered*)

Warum zentrieren?

Mit zentriertem Prädiktor gibt der Achsenabschnitt β_0 die Größe einer Person mit **durchschnittlichem** Gewicht an — sonst wäre es die (unsinnige) Größe bei Gewicht = 0. Vgl. Gelman et al. (2021), Kap. 10.4, 12.2.

Zentrieren — visualisiert



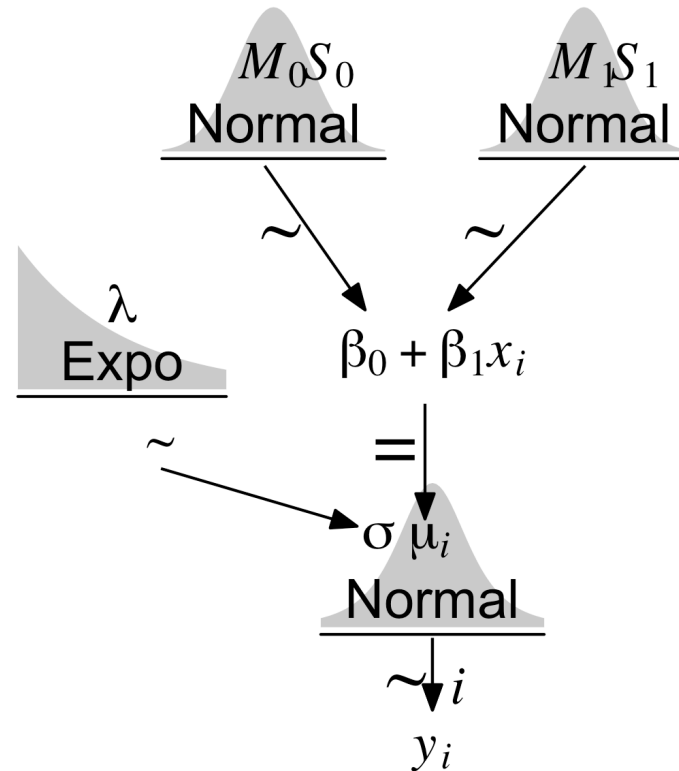
Die vertikalen gepunkteten Linien zeigen die Mittelwerte der Verteilungen.

Der Mittelwert von `weight_c` liegt danach bei 0.

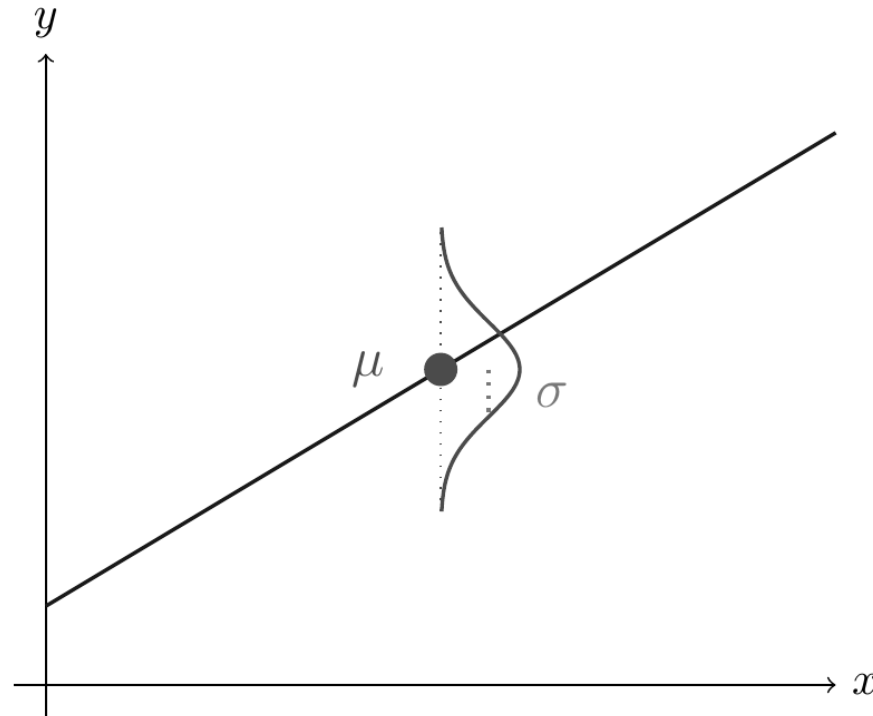
Modell
m_kung_gewicht_c:
zentrierter Prädiktor



Modellidee: drei Parameter zusammen



Vorhergesagte Größen als Normalverteilungen



Modelldefinition `m_kung_gewicht_c`

i Definition: Modelldefinition `m_kung_gewicht_c`

$\text{height}_i \sim \text{Normal}(\mu_i, \sigma)$

$\mu_i = \beta_0 + \beta_1 \cdot \text{weightcentered}_i$

$\beta_0 \sim \text{Normal}(178, 20)$

$\beta_1 \sim \text{Normal}(0, 10)$

$\sigma \sim \text{Exp}(0.1)$

Likelihood

Lineares Modell

Priori

Priori

Priori



Likelihood lesen

$$\text{height}_i \sim \text{Normal}(\mu_i, \sigma)$$

- Jetzt ein Index i an μ und h : nicht mehr **ein** Mittelwert, sondern **pro Beobachtung** ein μ_i

“Die Wahrscheinlichkeit, eine bestimmte Größe h bei Person i zu beobachten, gegeben μ und σ , ist normalverteilt.”



Das lineare Modell lesen

$$\mu_i = \beta_0 + \beta_1 \cdot \text{weightcentered}_i$$

- μ wird jetzt **deterministisch berechnet**, nicht geschätzt
- β_1 : Unterschied in **height** bei einer Einheit Unterschied in **weight** (Steigung)
- β_0 : Wert von μ , wenn **weight_c** = 0 (Achsenabschnitt)

“Der vorhergesagte mittlere Wert der Körpergröße, μ_i , berechnet sich als Summe von β_0 und $\beta_1 \cdot \text{weight}_i$.”



Priori-Verteilungen des Modells

$$\beta_0 \sim \text{Normal}(178, 20)$$

Priori Achsenabschnitt

$$\beta_1 \sim \text{Normal}(0, 10)$$

Priori Regressionsgewicht

$$\sigma \sim \text{Exp}(0.1)$$

Priori Sigma

- β_1 : schwach informativ — im Schnitt kein Zusammenhang, aber auch starke positive/negative Werte plausibel
- Zahl der Prioris = Zahl der Modellparameter



Die Post-Verteilung befragen

Ein “meinungsstärkeres” Modell

- Modell `m_kung_starkes_beta1`: andere Priori für β_1

$$\beta_1 \sim N(5, 3); \quad \beta_0 \sim N(178, 20); \quad \sigma \sim E(0.1)$$

- Im Schnitt: 5 cm Unterschied in Größe pro 1 kg Unterschied im Gewicht — deutlich “meinungsstärker” als zuvor

 Hinweis

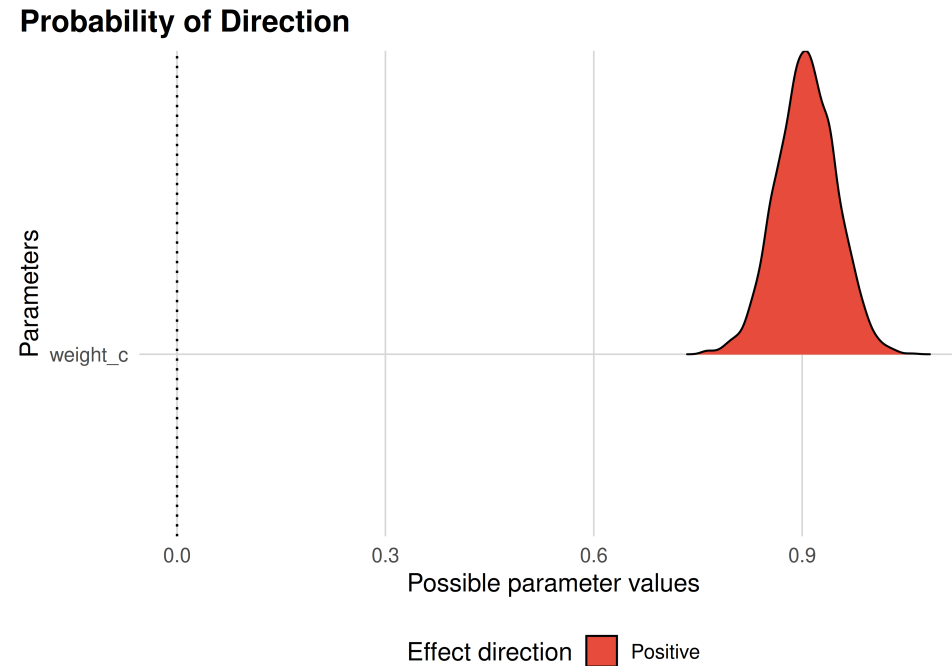
`seed` fixiert die Zufallszahlen für reproduzierbare Ergebnisse (Konvention hier: 42).

Definition: Effektwahrscheinlichkeit (pd)

Definition: Effektwahrscheinlichkeit

Die Kennzahl **pd** (*probability of direction*) gibt an, mit welcher Wahrscheinlichkeit ein Effekt positiv bzw. negativ ist (je nachdem, ob der Median des Effekts positiv oder negativ ist). Sie sagt nichts über die Effektstärke aus — nur ob der Effekt klar auf einer Seite der Null liegt. Grob vergleichbar mit dem p-Wert der Frequentistischen Statistik ([Makowski et al., 2019](#)).

Effektwahrscheinlichkeit visualisiert



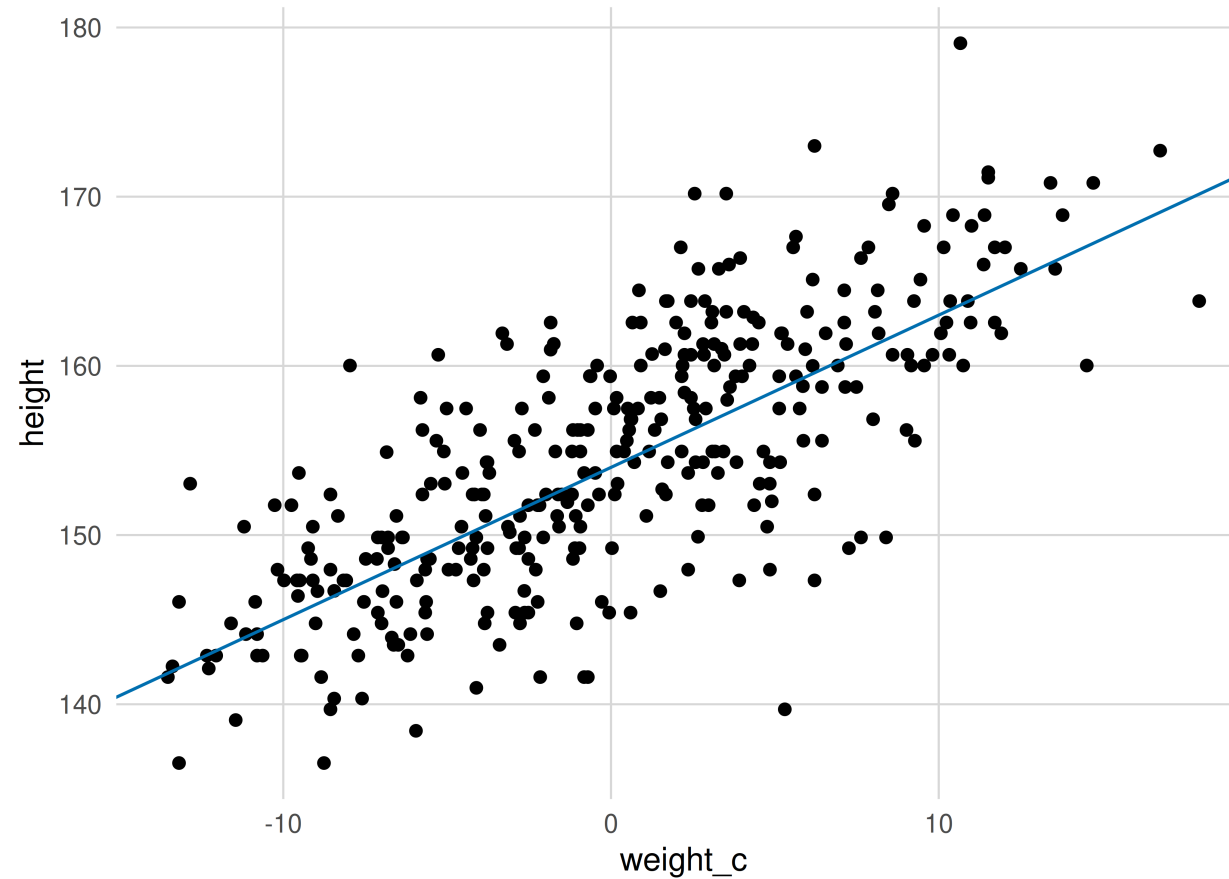
- [Rhat](#), [ESS](#): Diagnosekennzahlen der Stichprobenziehung — bei einfachen Modellen i. d. R. unproblematisch

Post-Verteilung befragen: die Bausteine

- Post-Verteilung = viele Stichproben zu β_0, β_1, σ
- `parameters(modell)`: Zusammenfassung der Post-Verteilung
- `as_tibble(modell)`: Rohe Stichproben als Tabelle
- Ergebnis für `m_kung_starkes_beta1`: β_1 (Steigung) liegt etwa zwischen 0.8 und 1.0 cm Größe pro kg Gewicht



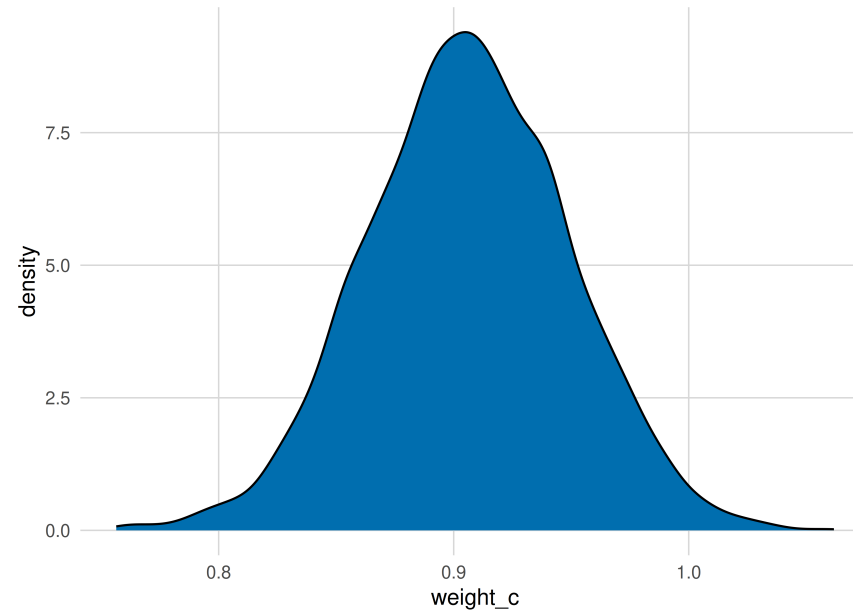
Die “mittlere” Regressionsgerade



Die Post-Verteilung befragen

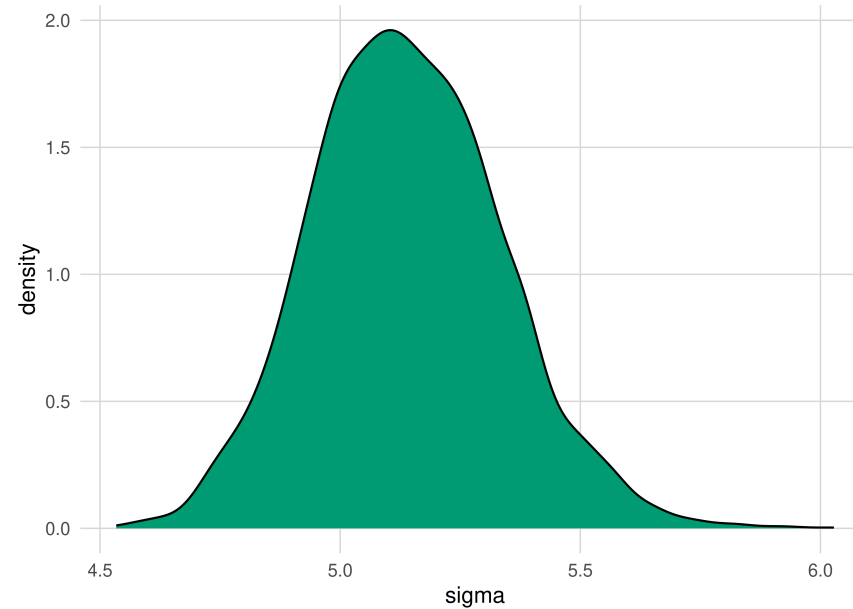


Post-Verteilung des Regressionsgewichts



- Mittelwert, Median und Modus (MAP) liegen eng beieinander

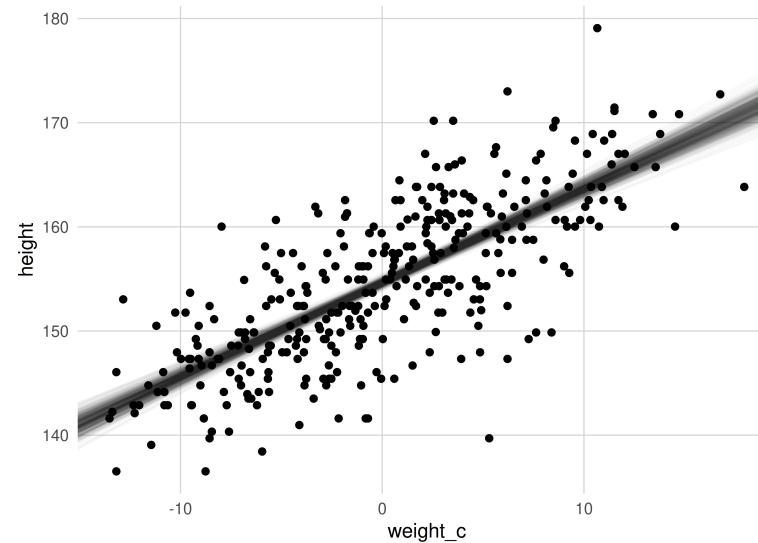
Post-Verteilung von Sigma



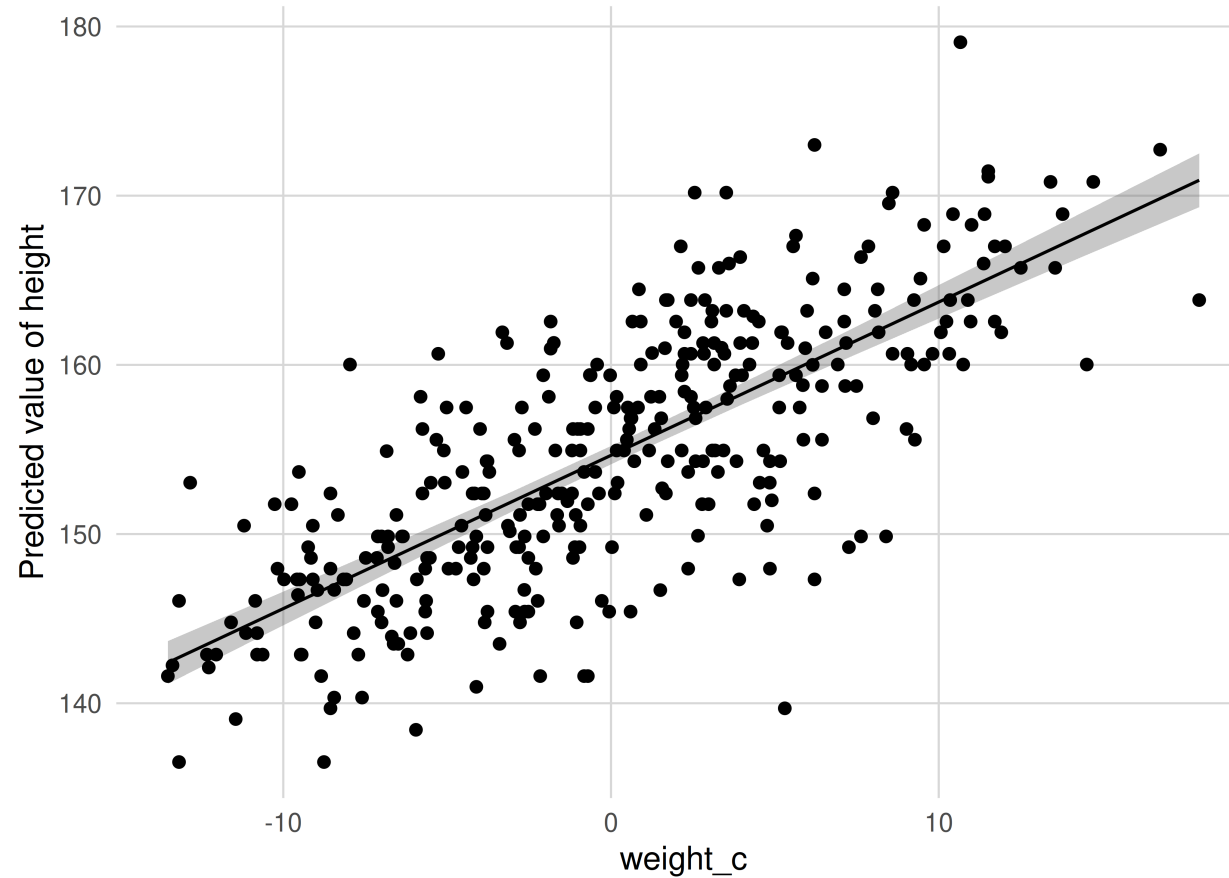
- Interessant: Die *Post*-Verteilung von σ ist normalverteilt, obwohl die *Apriori*-Verteilung exponentialverteilt war — die Daten haben den Prior “überstimmt”

Ungewissheit sichtbar machen

- Regressionsgeraden aus einzelnen Stichproben der Post-Verteilung übereinandergelegt
- 10 Stichproben → breites Band; 1000 Stichproben → dichtes, informatives Band um die wahrscheinlichste Gerade



Erwartungswert mit Unsicherheitsband



Die Post-Verteilung befragen



Fragen an die Post-Verteilung: Quantile

Typische Fragen an den (zentrierten) Achsenabschnitt:

- Welche mittlere Größe wird mit 50 %, 90 %, 95 % Wahrscheinlichkeit nicht überschritten?
- Welche mittlere Größe wird mit 95 % Wahrscheinlichkeit nicht unterschritten?
- Von wo bis wo reicht der innere 50 %-Schätzbereich (ETI)?

Hinweis

Statt eines 95 %-Intervalls wird teils ein 90 %- oder 89 %-Intervall empfohlen (bessere numerische Stabilität).

Fragen an die Post-Verteilung: Wahrscheinlichkeitsmassen

- Wie wahrscheinlich ist es, dass die mittlere Größe mindestens 155 cm beträgt?
- Wie wahrscheinlich ist es, dass die mittlere Größe höchstens 154.5 cm beträgt?

Beantwortbar durch einfaches Auszählen der Stichproben aus der Post-Verteilung (Anteil erfüllt Bedingung).

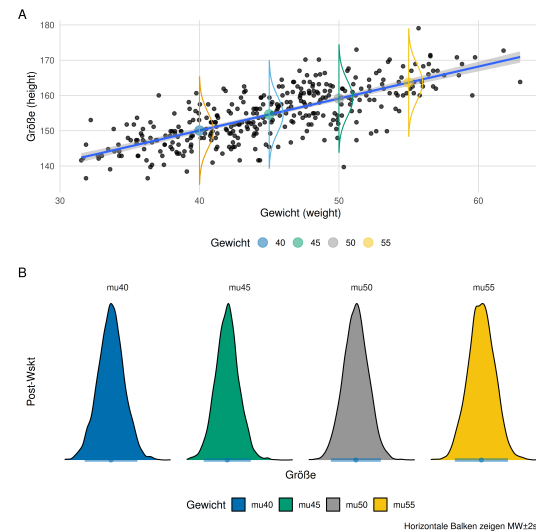


**Post-Verteilung bedingt
auf einen Prädiktorwert**



Post-Verteilung für jeden Gewichtswert

- Für **jeden** beliebigen Wert des Prädiktors lässt sich eine eigene Post-Verteilung von μ berechnen
- Beispiel: Wie groß ist eine Person (im Mittel) bei Gewicht 45 kg? Bei 50 kg?

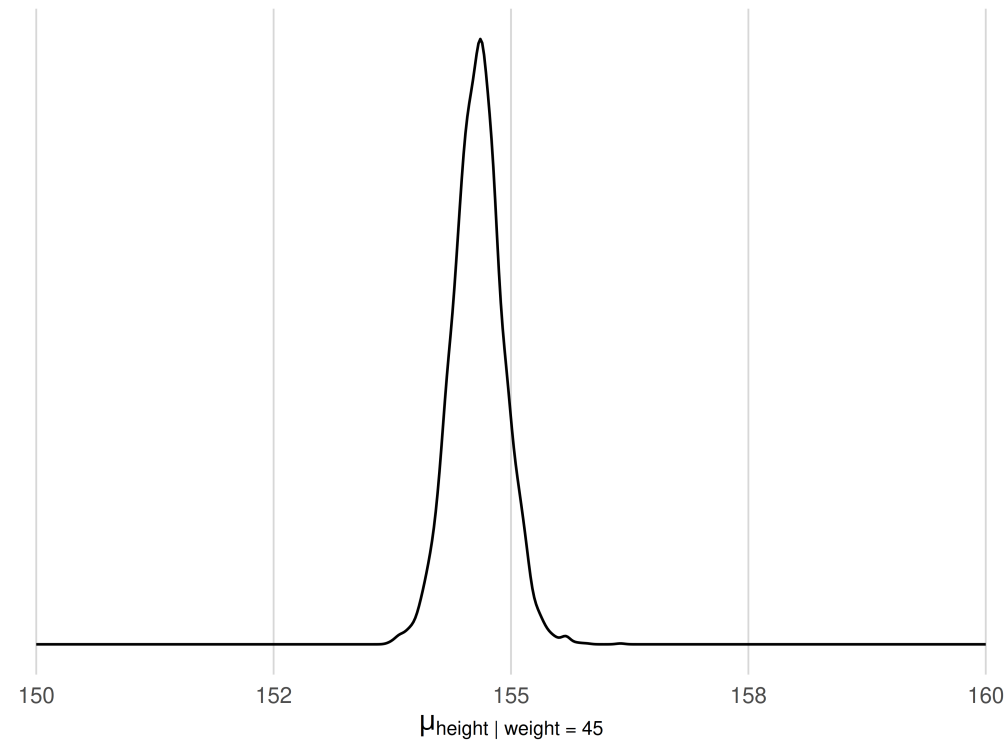


Post-Verteilung bedingt auf einen Prädiktorwert



Beispiel: μ bei Gewicht = 45 kg

- 45 kg entspricht dem Mittelwert von `weight` $\rightarrow$ `weight_c` = 0

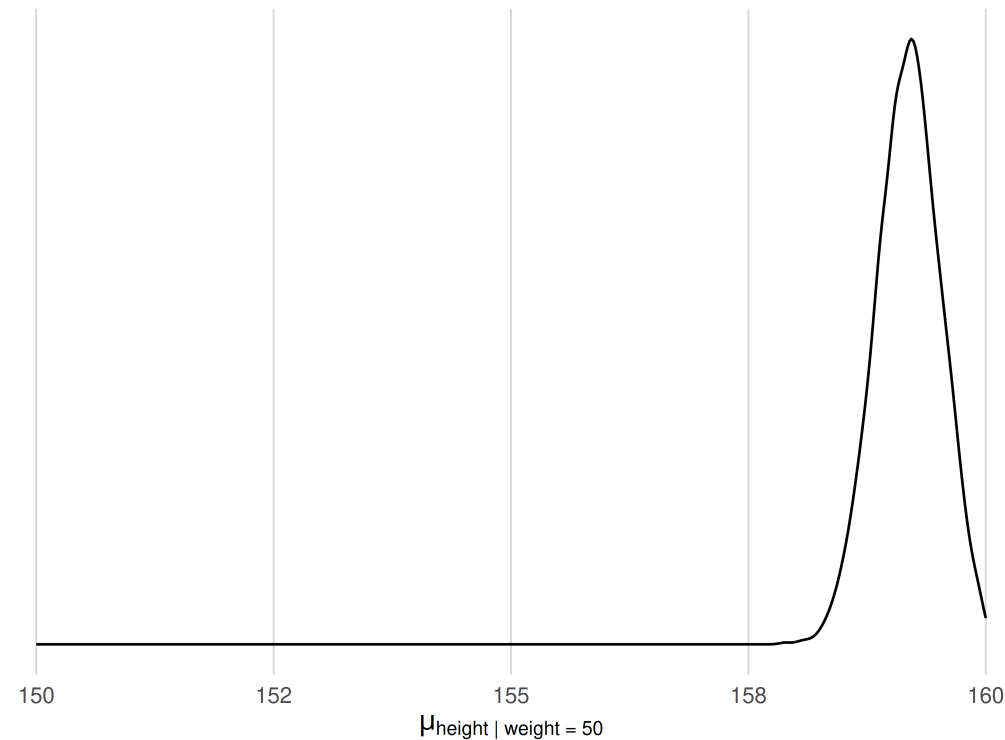


Post-Verteilung bedingt auf einen Prädiktorwert



Beispiel: μ bei Gewicht = 50 kg

- 50 kg = 5 über dem Mittelwert $\rightarrow$ `weight_c = 5`
- `mu_at_50 = (Intercept) + 5 * weight_c`



Post-Verteilung bedingt auf einen Prädiktorwert



Lagemaße und Streuungen der bedingten Post-Verteilung

- 90 %-ETI für $\mu|w = 50$: ca. 159 cm bis 160 cm
- 95 %-Quantil für $\mu|w = 45$: welche Größe wird mit 95 % nicht überschritten?

Dieselben Werkzeuge (Quantile, ETI/HDI, Auszählen) funktionieren für jede bedingte Post-Verteilung.



Fazit



Zusammenfassung

- Lineares Modell + Priori-Verteilungen $\rightarrow$ Post-Verteilung für β_0, β_1, σ
- Zentrieren des Prädiktors macht β_0 leichter interpretierbar
- Die Post-Verteilung lässt sich für **jeden** Prädiktorwert befragen: Lage, Streuung, Quantile, Wahrscheinlichkeitsmassen
- Mehr Information als Punktschätzer der Frequentistischen Regression



Vertiefende Literatur

- McElreath ([2020](#)), Kap. 4: vertiefte Darstellung linearer Bayes-Modelle
- Kurz ([2021](#)): Tidyverse-Umsetzung der Inhalte von McElreath ([2020](#))
- Gelman et al. ([2021](#)): Regressionsanalyse frequentistisch und Bayes



Vielen Dank!

Fragen?

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and Other Stories*. Cambridge University Press.

Kurz, S. (2021). *Statistical Rethinking with Brms, Ggplot2, and the Tidyverse: Second Edition*. <https://bookdown.org/content/4857/>

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10. <https://www.frontiersin.org/article/10.3389/fpsyg.2019.02767>

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2. Aufl.). Taylor and Francis, CRC Press.

