

# **Lineare Modelle**

**QM2, Thema 5**

**AWM, HS Ansbach**

# Gliederung

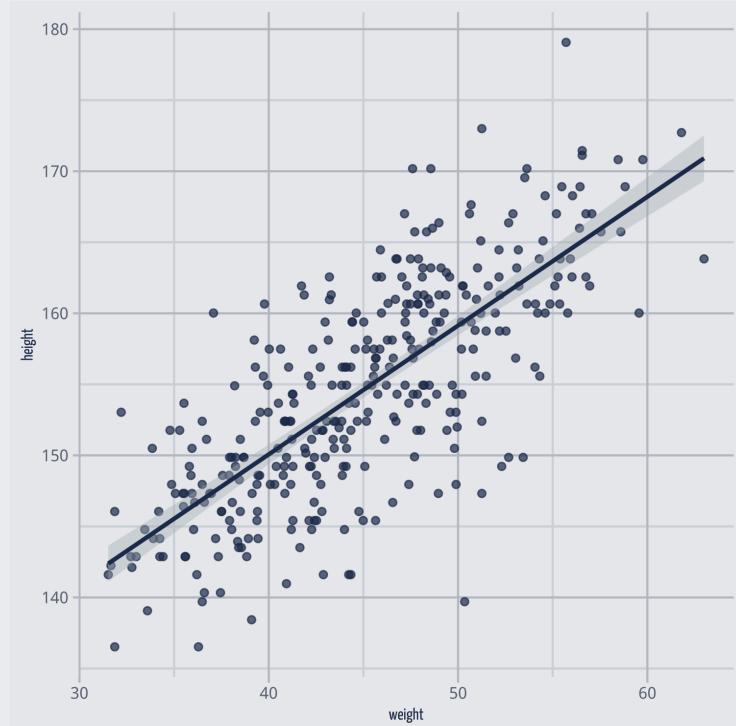
1. Teil 1: Die Post-Verteilung der Regression berechnen
2. Teil 2: Die Post-Verteilung befragen
3. Teil 3: Die PPV befragen
4. Hinweise

# Teil 1

## Post-Verteilung der Regression

# Einfache Regression

- Die (einfache) Regression prüft, inwieweit zwei Variablen,  $Y$  und  $X$  linear zusammenhängen.
- Je mehr sie zusammenhängen, desto besser kann man  $X$  nutzen, um  $Y$  vorherzusagen (und umgekehrt).
- Hängen  $X$  und  $Y$  zusammen, heißt das nicht (unbedingt), dass es einen *kausalen* Zusammenhang zwischen  $X$  und  $Y$  gibt.
- Linear bedeutet, der Zusammenhang ist additiv und konstant: wenn  $X$  um eine Einheit steigt, steigt  $Y$  immer um  $b$  Einheiten.



# Statistiken zum !Kung-Datensatz

## Datenquelle

```
library(tidyverse)
library(rstatix)
Kung_path <- "https://tinyurl.com/jr7ckxxj" # Datenquelle s.o.

d <- read_csv(Kung_path)

d2 <- d %>% filter(age > 18)

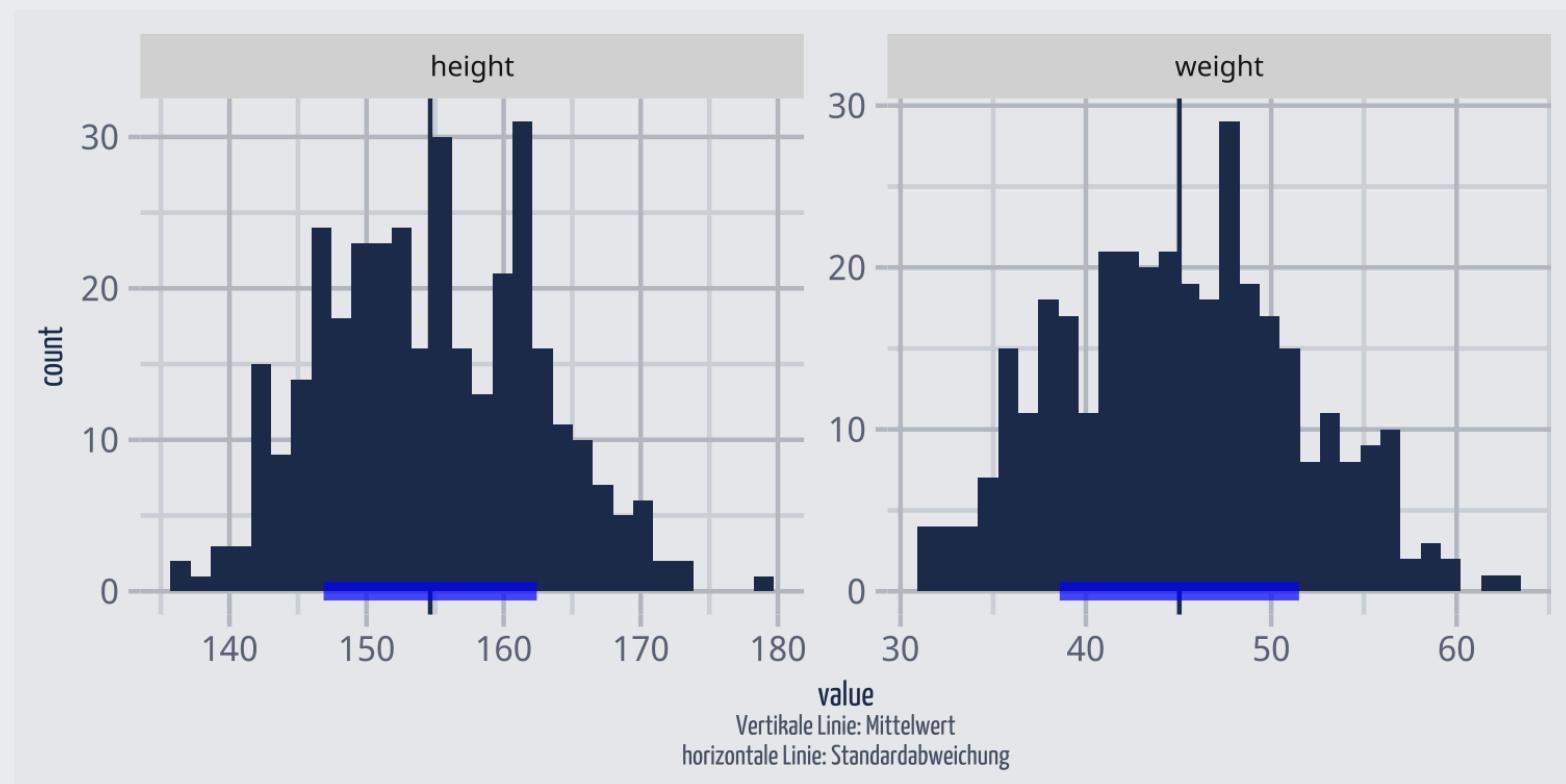
get_summary_stats(d2)
```

variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se	ci
age	346.0	19.0	88.0	40.0	29.0	51.0	22.0	16.3	41.5	15.8	0.8	1.7
height	346.0	136.5	179.1	154.3	148.6	160.7	12.1	8.5	154.6	7.8	0.4	0.8
male	346.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.5	0.5	0.0	0.1
weight	346.0	31.5	63.0	45.0	40.3	49.4	9.0	6.7	45.0	6.5	0.3	0.7

Das mittlere Körpergewicht (weight) liegt bei ca. 45kg (sd 7 kg).

# Visualisierung von weight und height

Explorative Datenanalyse (keine Inferenz auf Populationswerte, sondern auf die Stichprobe bezogen)



# Prädiktor zentrieren 1/2

- Zieht man von jedem Gewichtswert den Mittelwert ab, so bekommt man die Abweichung des Gewichts vom Mittelwert (Prädiktor "zentrieren").
- Wenn man den Prädiktor (`weight`) zentriert hat, ist der Achsenabschnitt,  $\alpha$ , einfacher zu verstehen.
- In einem Modell mit zentriertem Prädiktor (`weight`) gibt der Achsenabschnitt die Größe einer Person mit durchschnittlichem Gewicht an.
- Würde man `weight` nicht zentrieren, gibt der Achsenabschnitt die Größe einer Person mit `weight=0` an, was nicht wirklich sinnvoll zu interpretieren ist.

(Gelman, Hill, and Vehtari, 2021)

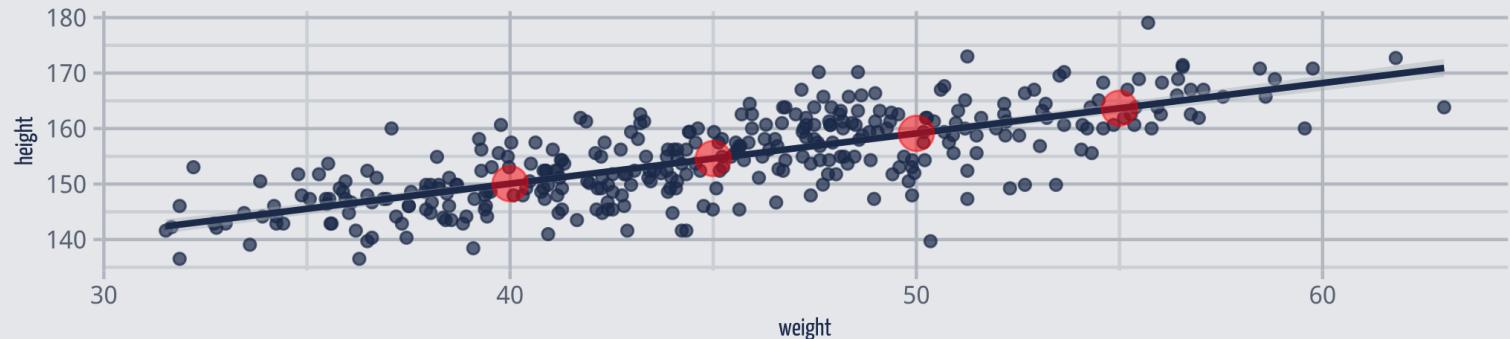
# Prädiktor zentrieren 2/2

```
d2 <-  
  d2 %>%  
  mutate(  
    weight_c = weight -  
    mean(weight))
```

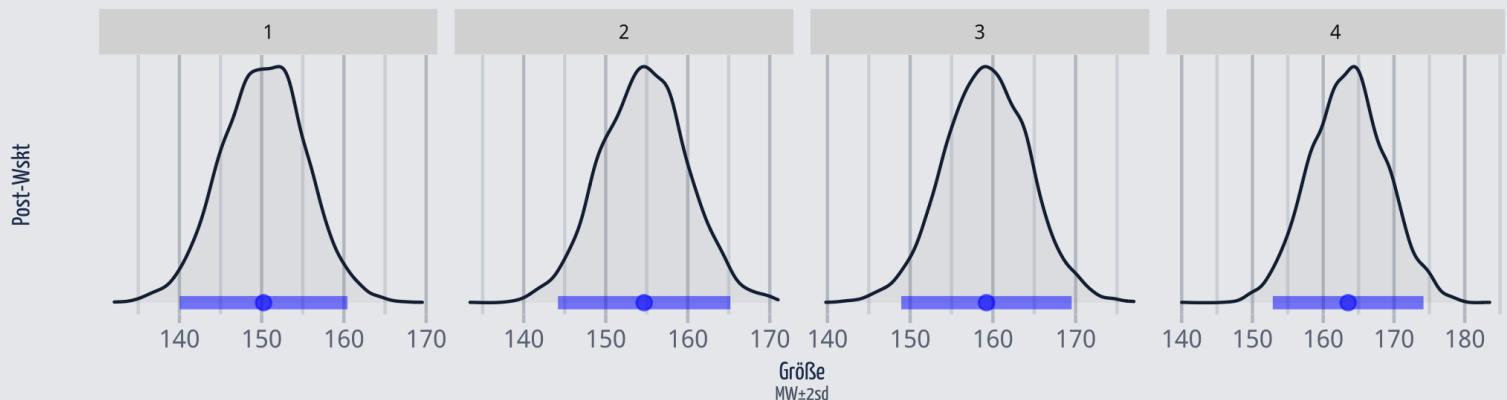
height	weight	age	male	weight_c
152	48	63	1	3
140	36	63	0	-9
137	32	65	0	-13

# Bei jedem Prädiktorwert eine Post-Verteilung für $\mu$

Für jeden Wert von X  
wird eine Post-Vert. berechnet



Post-Verteilungen an verschiedenen Werten von X



# Modelldefinition von m43

- Für jede Ausprägung des Prädiktors (weight),  $h_i$ , wird eine Post-Verteilung für die abhängige Variable (height) berechnet.
- Der Mittelwert  $\mu$  für jede Post-Verteilung ergibt sich aus dem **linearen Modell (unserer Regressionsformel)**.
- Die Post-Verteilung berechnet sich auf Basis der **Priori-Werte** und des **Likelihood** (Bayes-Formel).
- Wir brauchen **Priori-Werte** für die Steigung  $\beta$  und den Achsenabschnitt  $\alpha$  der Regressionsgeraden.
- Außerdem brauchen wir einen **Priori-Wert**, der die Streuung  $\sigma$  der Größe (height) angibt; dieser Wert wird als exponentialverteilt angenommen.
- Der **Likelihood** gibt an, wie wahrscheinlich ein Wert height ist, gegeben  $\mu$  und  $\sigma$ .

$\text{height}_i \sim \text{Normal}(\mu_i, \sigma)$	<b>Likelihood</b>
$\mu_i = \alpha + \beta \cdot \text{weight}_i$	<b>Lineares Modell</b>
$\alpha \sim \text{Normal}(178, 20)$	<b>Priori</b>
$\beta \sim \text{Normal}(0, 10)$	<b>Priori</b>
$\sigma \sim \text{Exp}(0.1)$	<b>Priori</b>

# Likelihood, m43

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

- Der Likelihood von m43 ist ähnlich zu den vorherigen Modellen (m41, m42).
- Nur gibt es jetzt ein kleines "Index-i" am  $\mu$  und am  $h$  (h wie heights).
- Es gibt jetzt nicht mehr nur einen Mittelwert  $\mu$ , sondern für jede Beobachtung (Zeile) einen Mittelwert  $\mu_i$ .
- Lies etwa so:

"Die Wahrscheinlichkeit, eine bestimmte Größe bei Person  $i$  zu beobachten, gegeben  $\mu$  und  $\sigma$  ist normalverteilt (mit Mittelwert  $\mu$  und Streuung  $\sigma$ )".

# Regressionsformel, m43

$$\mu_i = \alpha + \beta \cdot \text{weight}_i \quad \text{Lineares Modell}$$

- $\mu$  ist jetzt nicht mehr ein Parameter, der (stochastisch) geschätzt werden muss.  $\mu$  wird jetzt (deterministisch) berechnet. Gegeben  $\alpha$  und  $\beta$  ist  $\mu$  ohne Ungewissheit bekannt.
- $\text{weight}_i$  ist der Prädiktorwert (weight) der  $i$ ten Beobachtung, also einer !Kung-Person (Zeile  $i$  im Datensatz).
- Lies etwa so:

| "Der Mittelwert  $\mu_i$  der  $i$ ten Person berechnet sich als Summe von  $\alpha$  und  $\beta \cdot \text{weight}_i$ ".

- $\mu_i$  ist eine lineare Funktion von weight.
- $\beta$  gibt den Unterschied in height zweier Beobachtung an, die sich um eine Einheit in weight unterscheiden (Steigung der Regressionsgeraden).
- $\alpha$  gibt an, wie groß  $\mu$  ist, wenn weight Null ist.

# Priori-Werte der Regression, m43

$$\begin{array}{ll} \alpha \sim \text{Normal}(178, 20) & \text{Priori} \\ \beta \sim \text{Normal}(0, 10) & \text{Priori} \\ \sigma \sim \text{Exp}(0.1) & \text{Priori} \end{array}$$

- Parameter sind hypothetische Kreaturen: Man kann sie nicht beobachten, sie existieren nicht wirklich. Ihre Verteilungen nennt man Priori-Verteilungen.
- $\alpha$  wurde in m41 als  $\mu$  bezeichnet, da wir dort eine "Regression ohne Prädiktoren" berechnet haben.
- $\sigma$  ist uns schon als Parameter bekannt und behält seine Bedeutung.
- $\beta$  fasst unser Vorwissen, ob und wie sehr der Zusammenhang zwischen Gewicht und Größe positiv (gleichsinnig ist).
  - Moment. Dieser Prior,  $\beta$  erachtet positive und negative Zusammenhang als gleich wahrscheinlich?!
  - Sind wir wirklich indifferent, ob der Zusammenhang von Gewicht und Größe positiv oder negativ ist? **Nein, sind wir nicht.**

# Priori-Prädiktiv-Verteilung für m43

- Was denkt **wir** bzw. unser Golem *apriori* über den Zusammenhang von Größe und Gewicht?
- Um diese Frage zu beantworten ziehen wir Stichproben aus den Priori-Verteilungen des Modells, also für  $\alpha$ ,  $\beta$  und  $\sigma$ .

a	b	sigma
149.4	-6.8	9.3
153.9	12.8	16.1
170.0	-0.7	8.2
172.7	1.3	35.1
149.1	-5.6	5.0

Jede Zeile definiert eine Regressionsgerade.

# Prior-Prädiktiv-Simulation für m43 mit stan\_glm()

```
m43_prior_pred <-  
  stan_glm(height ~ weight_c,  
            prior = normal(0, 10),  
            prior_intercept = normal(178, 20), # mu  
            prior_aux = exponential(0.1), # sigma  
            refresh = FALSE,  
            prior_PD = TRUE, # DIESER Schalter macht  
            data = d2)  
  
m43_prior_pred_draws <-  
  m43_prior_pred %>%  
  as_tibble() %>%  
  rename(a = `(Intercept)` ,  
         b = weight_c) %>%  
  slice_sample(n = 50)
```

]

# Visualisieren der Prior-Prädiktiv-Verteilung

```
d2 %>% ggplot() +  
  geom_point(aes(x = weight_c, y = height)) +  
  geom_abline(data = m43_prior_pred_draws,  
    aes(intercept = a, slope = b), color = "skyblue", size = 0.2) +  
  scale_y_continuous(limits = c(0, 500)) +  
  geom_hline(yintercept = 272, size = .5) +  
  geom_hline(yintercept = 0, linetype = "dashed")
```

😱 Einige dieser Regressionsgeraden sind unsinnig!

# Ein positiver Wert für $\beta$ ist plausibler

Oh no

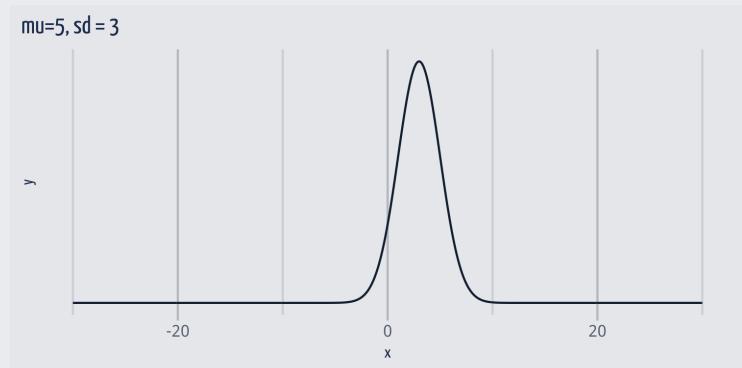
Eine Normalverteilung mit viel Streuung:



👎  $\beta = -20$  wäre mit diesem Prior gut möglich: Pro kg Gewicht sind Menschen im Schnitt 20cm kleiner, laut dem Modell. Quatsch.

Oh yes

Wir bräuchten eher so eine Verteilung, mit mehr Masse auf der positiven Seite ( $x > 0$ ):



👍 Vermutlich besser: Ein Großteil der Wahrscheinlichkeitsmasse ist  $X > 0$ . Allerdings gibt's keine Gewähr, dass unser Prior "richtig" ist.

# Priori-Prädiktiv-Simulation, 2. Versuch

a	b	s
159.9	-1.3	34.9
162.5	-1.1	25.6
173.9	-0.4	21.0
153.2	-0.7	14.4
146.2	-1.6	13.5

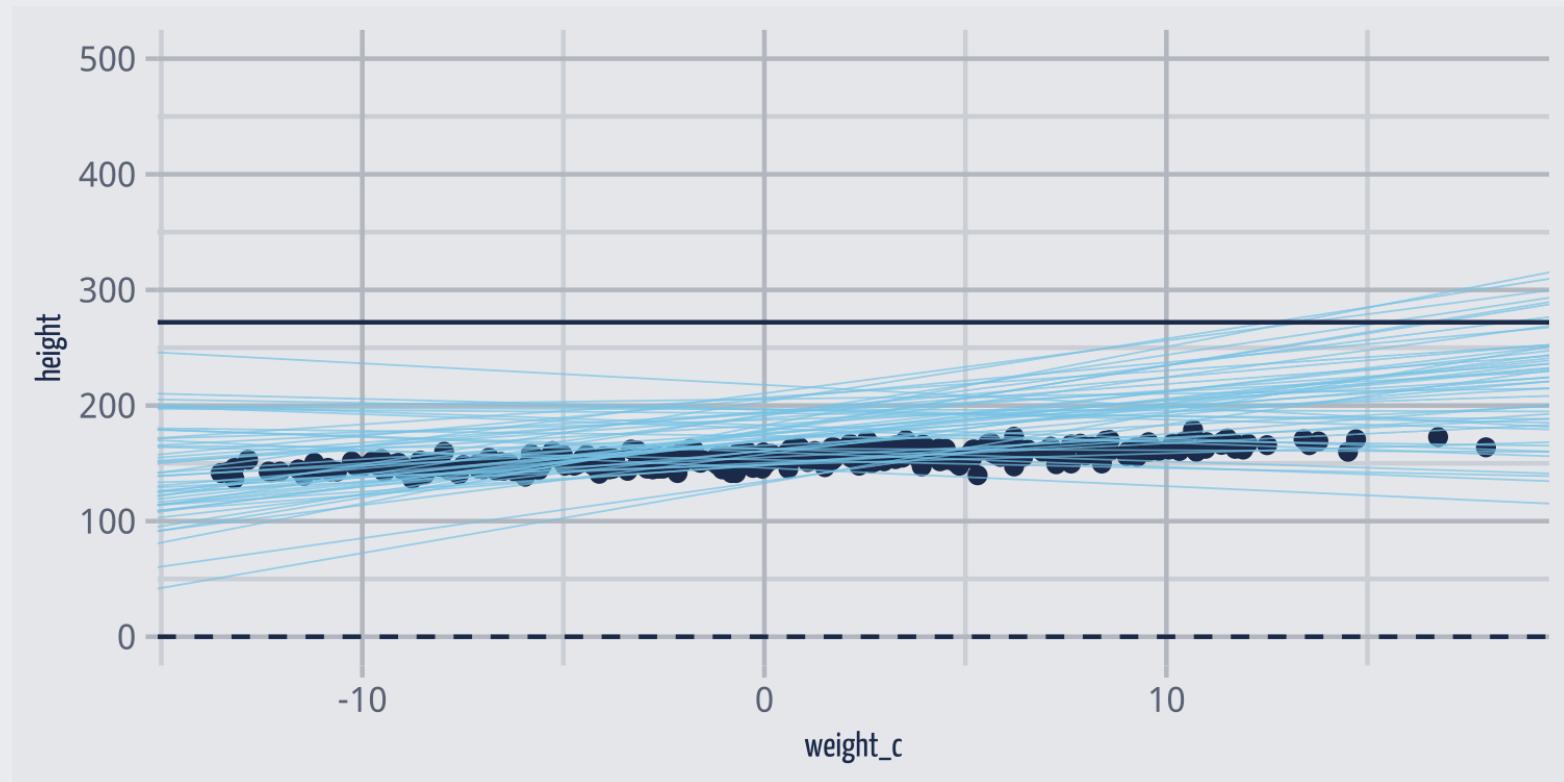
```
m43a_prior_pred <-
  stan_glm(
    height ~ weight_c,
    prior = normal(2, 2), # Regressionsgewicht
    prior_intercept = normal(178, 20), # mu
    prior_aux = exponential(0.1), # sigma
    refresh = FALSE,
    # Schalter für Prior-Pred-Verteilung:
    prior_PD = TRUE,
    data = d2)

m43a_prior_pred_draws <-
  m43a_prior_pred %>%
  as_tibble() %>%
  # Spaltennamen kürzen:
  rename(a = `Intercept`) %>%
  rename(b = weight_c,
         s = sigma)
```

Das Argument `prior_PD = TRUE` sorgt dafür, dass keine Posteriori-Verteilung, sondern eine Prior-Prädiktiv-Verteilung berechnet wird.

# Visualisieren der Prior-Prädiktiv-Verteilung, m43a

Unsere Priori-Werte scheinen einigermaßen vernünftige Vorhersagen zu tätigen. Allerdings erwartet unser Golem einige Riesen.



Die durchgezogene horizontale Linie gibt die Größe des [größten Menschen, Robert Pershing Wadlow](#), an.

# Moment, kann hier jeder machen, was er will?

Es doch den einen, richtigen, objektiven Priori-Wert geben?!

Kann denn jeder hier machen, was er will?! Wo kommen wir da hin?!

This is a mistake. There is no more a uniquely correct prior than there is a uniquely correct likelihood. Statistical models are machines for inference. Many machines will work, but some work better than others. Priors can be wrong, but only in the same sense that a kind of hammer can be wrong for building a table. [McElreath \(2020\)](#), p. 96.

# Hier ist unser Modell, m43a

$$\begin{aligned} \text{height}_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta \cdot \text{weight}_i \\ \alpha &\sim \text{Normal}(178, 20) \\ \beta &\sim \text{Normal}(5, 3) \\ \sigma &\sim \text{Exp}(0.1) \end{aligned}$$

```
# Zufallszahlen festlegen:  
set.seed(42)  
# Posteriori-Vert. berechnen:  
m43a <-  
  stan_glm(  
    height ~ weight_c, # Regressionsformel  
    prior = normal(5, 3), # beta  
    prior_intercept = normal(178, 20), # mu  
    prior_aux = exponential(0.1), # sigma  
    refresh = 0, # zeig mir keine Details  
    data = d2)
```

# Eine Zusammenfassung der Posteriori-Verteilung für m43a

```
## stan_glm
##   family:      gaussian [identity]
##   formula:     height ~ weight_c
##   observations: 346
##   predictors:  2
## -----
##           Median MAD_SD
## (Intercept) 154.6    0.3
## weight_c     0.9    0.0
##
## Auxiliary parameter(s):
##           Median MAD_SD
## sigma 5.1    0.2
##
## -----
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

## Teil 2

Die Post-Verteilung befragen

# Mittelwerte von $\alpha$ und $\beta$ aus der Post-Verteilung

```
post_m43a <-  
  as_tibble(m43a)
```

Die ersten paar Zeilen:

id	(Intercept)	weight_c	sigma
1	154.8	0.9	4.9
2	154.7	0.8	4.8
3	154.9	1.0	5.1

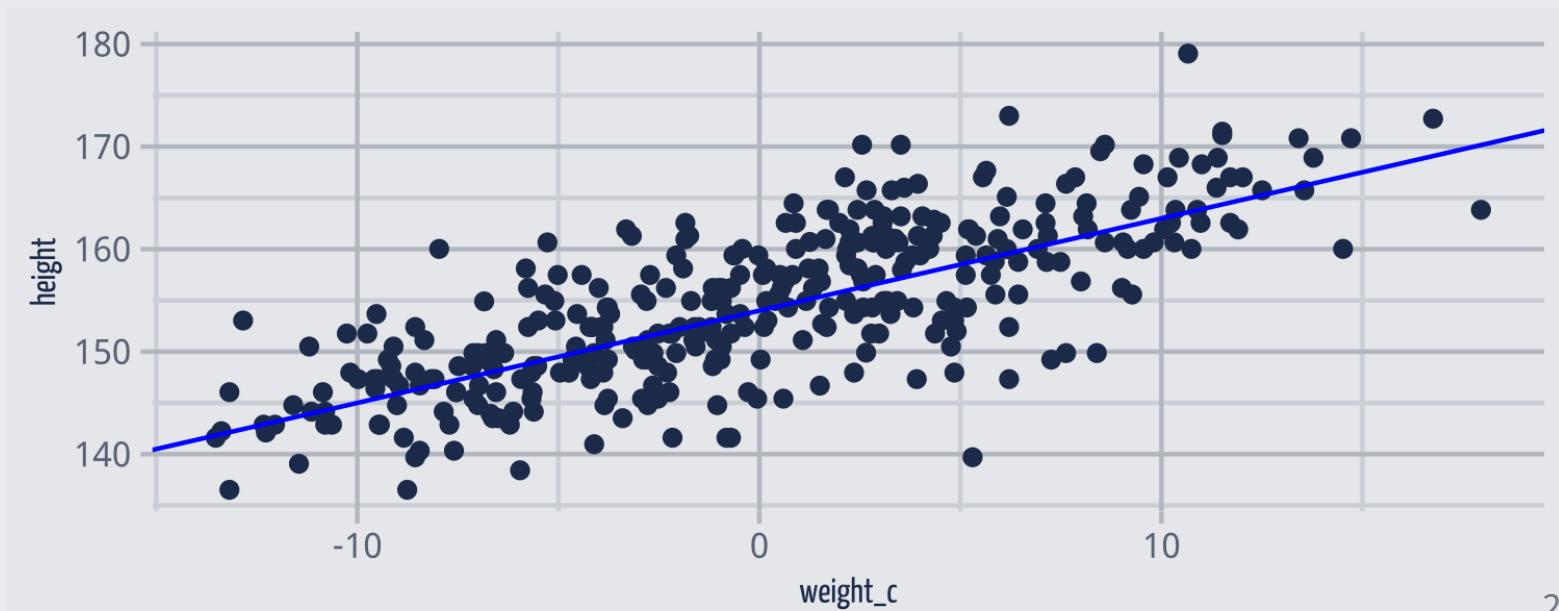
```
names(post_m43a) <-  
  c("a", "b", "sigma")  
  
post_m43a_summary <-  
  post_m43a %>%  
  summarise(  
    a_mean = mean(a),  
    b_mean = mean(b),  
    s_mean = mean(sigma))
```

a_mean	b_mean	s_mean
154.7	0.9	5.1

# Visualisieren der "mittleren" Regressionengeraden

a_mean	b_mean	s_mean
154.7	0.9	5.1

```
d2 %>%
  ggplot() +
  aes(x = weight_c, y = height) +
  geom_point() +
  geom_abline(
    slope = 0.9,
    intercept = 154,
    color = "blue")
```



# Zentrale Statistiken zu den Parametern

In diesem Modell gibt es drei Parameter:  $\mu, \beta, \sigma$ .

## Mittelwerte

- Mittlere Größe?
- Schätzwert für den Zusammenhang von Gewicht und Größe?
- Schätzwert für Ungewissheit in der Schätzung der Größe?

```
post_m43a_summary
```

```
## # A tibble: 1 × 3
##   a_mean b_mean s_mean
##     <dbl>  <dbl>  <dbl>
## 1 155.    0.908   5.14
```

## Streuungen

- Wie unsicher sind wir uns in den Schätzungen der Parameter?

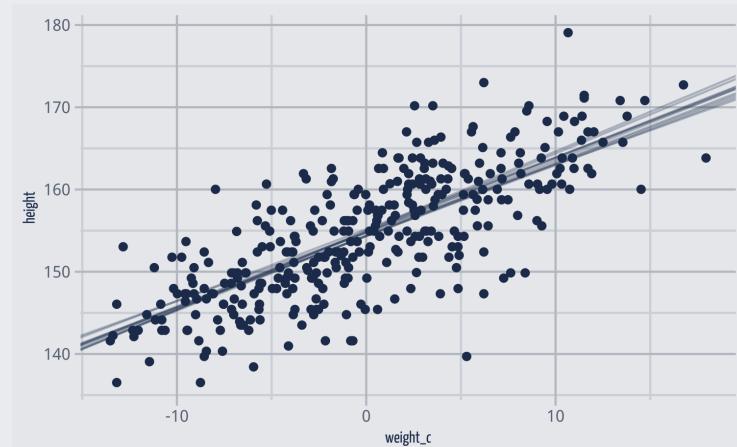
```
post_m43a_summary2 <-
  post_m43a %>%
  summarise(
    a_sd = sd(a),
    b_sd = sd(b),
    s_sd = sd(sigma))
```

a_sd	b_sd	s_sd
0.28	0.04	0.19

# Ungewissheit von $\alpha$ und $\beta$ aus der Post-Verteilung

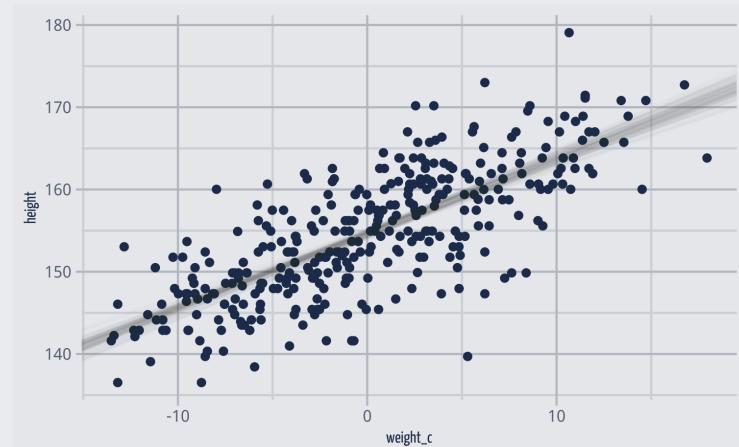
Die ersten 10 Stichproben

```
d2 %>%
  ggplot(aes(x = weight_c,
             y = height)) +
  geom_point() +
  geom_abline(
    data = post_m43a %>% slice_he
    aes(slope = b,
        intercept = a),
    alpha = .3)
```



Die ersten 100 Stichproben

```
d2 %>%
  ggplot(aes(x = weight_c,
             y = height)) +
  geom_point() +
  geom_abline(
    data = post_m43a %>% slice_he
    aes(slope = b,
        intercept = a),
    alpha = .02)
```



# Fragen zu Quantilen des Achsenabschnitts

Bei einem zentrierten Prädiktor misst der Achsenabschnitt die mittlere Größe.

- Welche mittlere Größe mit zu 50%, 90% Wskt. nicht überschritten?
- Welche mittlere Größe mit zu 95% Wskt. nicht unterschritten?
- Von wo bis wo reicht der innere 50%-Schäzbereich der mittleren Größe?

```
## # A tibble: 1 × 3
##      q_50    q_90    q_05
##      <dbl>   <dbl>   <dbl>
## 1 155.    155.    154.

## # A tibble: 2 × 1
##      pi_50
##      <dbl>
## 1 154.
## 2 155.
```

```
post_m43a %>%
  summarise(
    q_50 =
      quantile(a, prob = .5),
    q_90 =
      quantile(a, prob = .9),
    q_05 =
      quantile(a, prob = .05))

post_m43a %>%
  summarise(
    pi_50 =
      quantile(a,
                prob = c(.25, .75)))
```

# Fragen zu Wahrscheinlichkeitsmassen des Achsenabschnitts

Bei einem zentrierten Prädiktor misst der Achsenabschnitt die mittlere Größe.

- Wie wahrscheinlich ist es, dass die mittlere Größe bei mind. 155 cm liegt?

```
post_m43a %>%
  count(gross = a >= 155) %>%
  mutate(prop = n / sum(n))
```

```
## # A tibble: 2 × 3
##   gross     n   prop
##   <lgl> <int> <dbl>
## 1 FALSE    3574 0.894
## 2 TRUE      426  0.106
```

Die Wahrscheinlichkeit beträgt 0.11.

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

```
post_m43a %>%
  count(klein = (a <= 154.5)) %>%
  mutate(prop = n / sum(n))
```

```
## # A tibble: 2 × 3
##   klein     n   prop
##   <lgl> <int> <dbl>
## 1 FALSE    2810 0.702
## 2 TRUE     1190 0.298
```

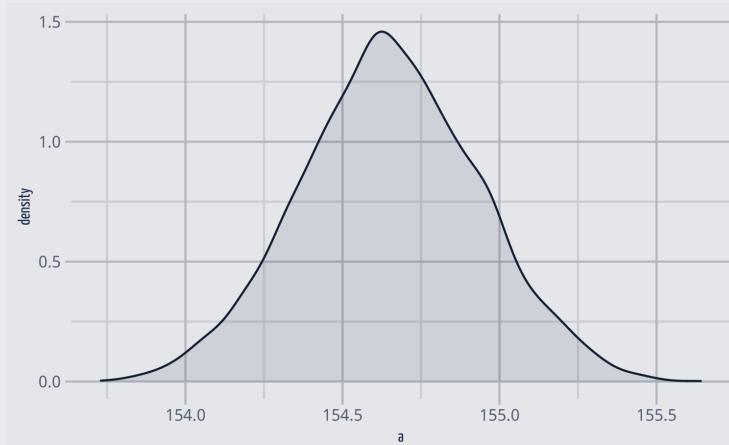
Die Wahrscheinlichkeit beträgt 0.3.

# Ungewissheit von Achsenabschnitt und Steigung

... als Histogramme visualisiert

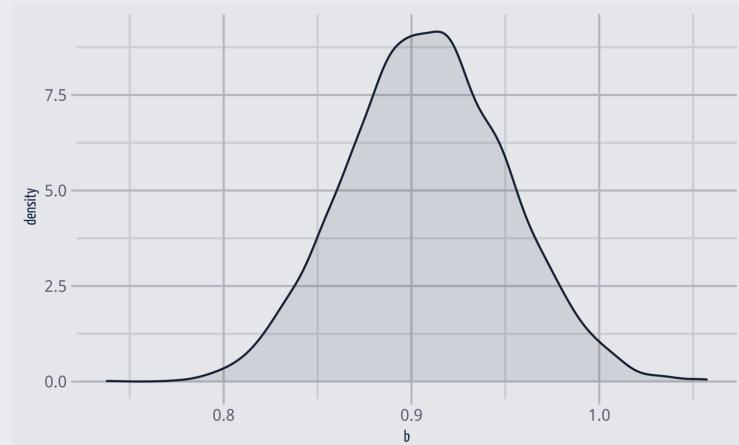
## Achsenabschnitt

```
post_m43a %>%
  ggplot(aes(x = a)) +
  geom_density()
```



## Regressionsgewicht (Steigung)

```
post_m43a %>%
  ggplot(aes(x = b)) +
  geom_density()
```



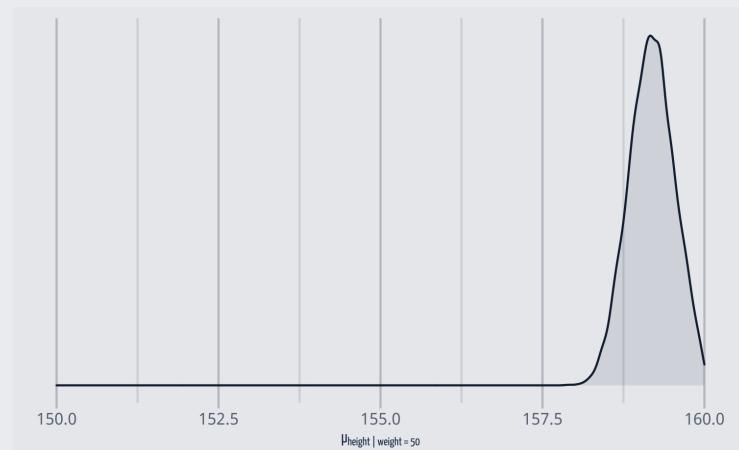
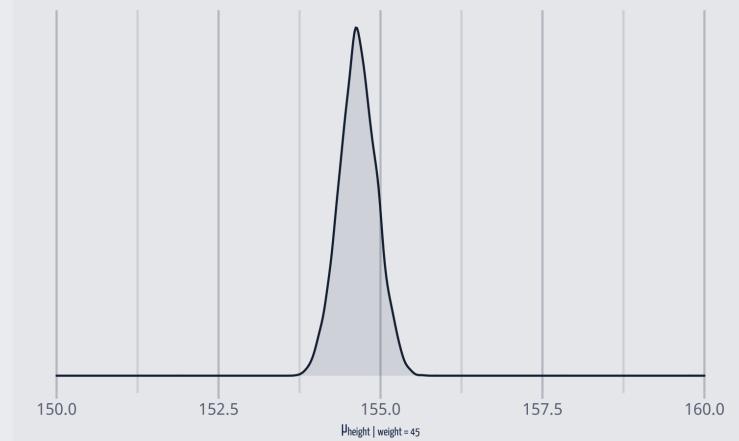
# Ungewissheit für $\mu | \text{weight} = 45, 50$

- 50 kg ist 5 kg über dem MW
- b ist zentriert: b=0 ist MW von weight

```
mu_at_45_50 <-
  post_m43a %>%
  mutate(mu_at_45 = a,
        mu_at_50 = a + b * 5)
```

```
mu_at_45_50 %>%
  ggplot(aes(x = mu_at_45)) +
  geom_density()
```

```
mu_at_45_50 %>%
  ggplot(aes(x = mu_at_50)) +
  geom_density()
```



# Wie groß ist ein !Kung mit 50kg Gewicht im Mittel?

$$\mu|w = 50$$

```
mu_at_45_50 %>%
  summarise(pi = quantile(mu_at_50, prob = c(0.5, .9)))
```

```
## # A tibble: 2 × 1
##       pi
##   <dbl>
## 1 159.
## 2 160.
```

Die mittlere Größe - gegeben  $w = 50$  - liegt mit 90% Wahrscheinlichkeit zwischen den beiden Werten.

Welche mittlere Größe wird mit 95% Wahrscheinlichkeit nicht überschritten, wenn die Person 45kg wiegt?

```
## # A tibble: 1 × 1
##       q_95
##   <dbl>
## 1 155.
```

# Teil 3

Die PPV befragen

# Perzentil-Intervalle für verschiedenen Prädiktor-Werte

Wir erstellen uns eine Sequenz an Prädiktorwerten, die uns interessieren:

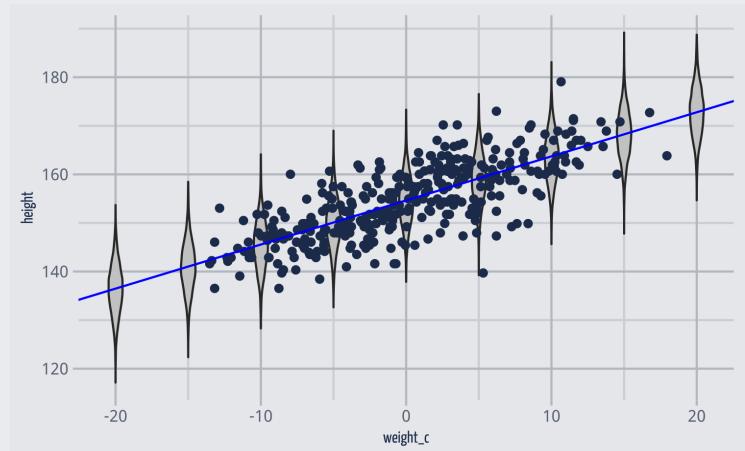
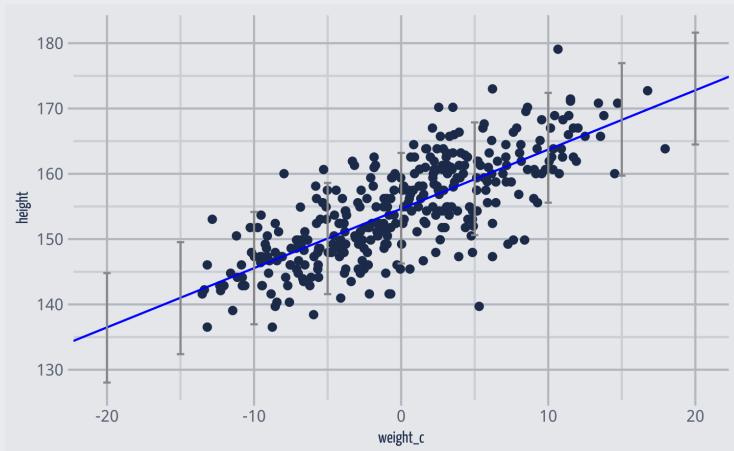
Für diese Werte lassen wir uns dann die Perzentil-Intervalle ausgeben:

```
mus <-  
  predictive_interval(  
    m43a,  
    newdata = weight_df) %>%  
    as_tibble() %>%  
    bind_cols(weight_df)
```

weight_c	5%	95%
-20.0	128.0	144.8
-15.0	132.4	149.5
-10.0	137.0	154.2
-5.0	141.6	158.6
0.0	146.3	163.2
5.0	150.6	167.9
10.0	155.6	172.4
15.0	159.7	176.9
20.0	164.5	181.6

Um die Perzentilintervalle zu erstellen, wird für jeden Prädiktorwert eine Posteriori-Verteilung erstellt und das 5%- sowie 95%-Quantil berechnet.

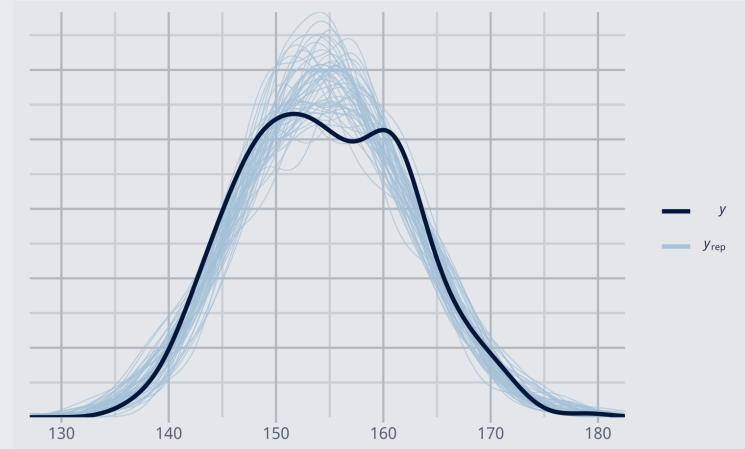
# Perzentilintervalle für verschiedenen Prädiktorwerte visualisiert



# Die PPV visualisiert

Vergleichen wir die echten Werte für height,  $h$ , mit den von der PPV simulierten Werten für height,  $h_{sim}$ .

```
library(bayesplot)
h <- d2$height
h_sim <-
  posterior_predict(m43a,
                     draws = 50)
ppc_dens_overlay(
  h, h_sim)
```



Die zwei Gipfel hat unser Modell nicht mitgekriegt, ansonsten decken sich die Vorhersagen der PPV gut mit den echten Daten.

# PPV plotten, von Hand

```
set.seed(42)
ppv_m43a <- posterior_predict(
  m43a,
  newdata = weight_df,
  draws = 100) %>%
  as_tibble() %>%
  pivot_longer(
    cols = everything(),
    names_to = "weight_condition",
    values_to = "height")
```

```
ppv_m43a %>%
  ggplot(aes(x = height)) +
  geom_density()
```

# Fragen an die PPV

- Wie groß sind die !Kung im Schnitt?
- Welche Größe wird von 90% der Personen nicht überschritten?
- Wie groß sind die 10% kleinsten?

```
ppv_m43a %>%
  summarise(
    q_10 = quantile(
      height, prob = .1),
    height_mean = mean(height),
    q_50 = quantile(
      height, prob = .5),
    q_90 = quantile(
      height, prob = .9)
  )
```

```
## # A tibble: 1 × 4
##   q_10  height_mean  q_50  q_90
##   <dbl>       <dbl> <dbl> <dbl>
## 1 138.        154. 154. 172.
```

- Was ist der 50% Bereich der Körpergröße?

```
ppv_m43a %>%
  summarise(
    pi_50 = quantile(
      height,
      prob = c(.25, .75))
  )
```

```
## # A tibble: 2 × 1
##   pi_50
##   <dbl>
## 1 144.
## 2 165.
```

# Hinweise

# Zu diesem Skript

- Dieses Skript bezieht sich auf folgende Lehrbücher:
  - Rethink, Kap. 4.4, ROS, Kap. 9.2
- Dieses Skript wurde erstellt am 2021-11-05 12:34:58
- Lizenz: CC-BY
- Autor ist Sebastian Sauer.
- Um diese HTML-Folien korrekt darzustellen, ist eine Internet-Verbindung nötig.
- Mit der Taste ? bekommt man eine Hilfe über Shortcuts.
- Wenn Sie die Endung .html in der URL mit .pdf ersetzen, bekommen Sie die PDF-Version der Datei. Wenn Sie mit .Rmd ersetzen, den Quellcode.
- Eine PDF-Version kann erzeugt werden, indem man im Chrome-Browser drückt (Drucken als PDF).

# Literatur

Gelman, A., J. Hill, and A. Vehtari (2021). *Regression and other stories*. Analytical methods for social research. Cambridge University Press.

McElreath, R. (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2nd ed. CRC texts in statistical science. Taylor and Francis, CRC Press.