

# Einführung in die Bayes Statistik- Slides Begleitung zum Online Kurs

## Modellierung und Posteriori

Martje Rave

Sommersemester 2025

Diese Folien sind **NUR** Kurs begleitend und können noch viele Fehler beinhalten. Für die Klausurvorbereitung benutzt bitte den Online-Kurs und die Übungsblätter.  
Bitte gebt mir Bescheid, wenn ihr Fehler findet.

## Normalverteilungsmodell mit einem unbekannten Parametern

- Wir nehmen an, dass das Gewicht eines Schokoladenhasen normalverteilt ist. Laut Hersteller liegt der Erwartungswert des Gewichts bei  $\mu = 101$  Gramm, die Standardabweichung bei der Herstellung bei  $\sigma^2 = 0.5$ .
- Wir glauben die Angabe  $\mu = 101$  nicht und behandeln den Parameter im Folgenden als unbekannt. Für die Bayesianische Analyse brauchen wir:
  - Die Datendichte (Likelihood)
  - Die Priori für den unbekannten Parameter  $\mu$

- Wir haben also  $n = 10$  unabhängig normalverteilte Beobachtungen

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, n.$$

## Datendichte- Likelihood

- Auf Grund der Unabhängigkeit der Beobachtungen ist die gemeinsame Datendichte das Produkt der einzelnen, also mit  $\mathbf{x} = (x_1, \dots, x_n)$

$$f(\mathbf{x}|\mu) = \prod_{i=1}^n f(x_i|\mu) \tag{1}$$

$$= \prod_{i=1}^n \left( \frac{1}{\sqrt{2\sigma^2\pi}} \right) \exp\left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \tag{2}$$

$$= \frac{1}{(\sqrt{2\sigma^2\pi})^n} \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \tag{3}$$

Die Kern Dichte damit:  $\exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right)$

# Konjugierte Priori

- Nun brauchen wir eine Priori für  $\mu$ .
- Wir kennen bisher den Ansatz der *konjugierten Priori*. Das heißt, die Posterioriverteilung ist die selbe wie die Priori-Verteilung, aber mit anderen Parametern. Dafür muss der Kern der Priori zum Kern der Datendichte passen.
- Der Kern der Datendichte für eine Beobachtung ist  $\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right)$ . Hält man im Kern  $x$  fest, handelt es sich wieder um den Kern einer Normalverteilung für  $\mu$ .

$$\mu \sim N(\mu_0, \sigma_0^2),$$

Dichte:

$$p(\mu) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2\right)$$

- $\mu_0$ : Erwartungswert
- $\sigma_0^2$ : Varianz (beides Priori-Parameter)

Posteriordichte ergibt sich (bis auf Konstante) durch Produkt von Likelihood und Priori:

$$\begin{aligned} p(\mu|x) &\propto f(x|\mu) \cdot p(\mu) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \cdot \exp\left(-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2} \left[ \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 + \frac{1}{\sigma_0^2} (\mu - \mu_0)^2 \right]\right) \end{aligned}$$

$$p(\mu|x) \propto \exp\left(-\frac{1}{2\sigma^2}(\mu - \hat{\mu})^2\right)$$

mit:

$$\hat{\mu} = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)$$

und

$$\hat{\sigma}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

Zur Vereinfachung:

$$\begin{aligned}\bar{m} &= \frac{n}{\sigma^2} \bar{x} + \frac{1}{\sigma_0^2} \mu_0 \\ \tau &= \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \\ \Rightarrow \mu|x &\sim \mathcal{N}\left(\frac{\bar{m}}{\tau}, \frac{1}{\tau}\right)\end{aligned}$$

Dies ist die Posteriorverteilung mit:

- Erwartungswert:  $\frac{\bar{m}}{\tau}$
- Varianz:  $\frac{1}{\tau}$

Wir wollen versuchen, eine **nicht-informative Priori** herzuleiten.

Dazu betrachten wir wieder die Posteriorparameter:

$$\bar{m} = \frac{\mu_0}{\sigma_0^2} + \frac{\sum x_i}{\sigma^2}, \quad \tau = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

- Entscheidend ist  $\sigma_0^2$ : Für  $\sigma_0^2 \rightarrow \infty$  verschwindet der Einfluss der Priori.
- Dann ist die Posteriorverteilung vollständig datengetrieben.
- Für  $\sigma_0^2 \rightarrow \infty$  entfällt die Priori:  $\Rightarrow$  **nichtinformative Priori**.

Zum Vergleich: **Jeffreys' Prior** ergibt sich als:

$$p(\mu) \propto \text{const.}$$

Dies entspricht:

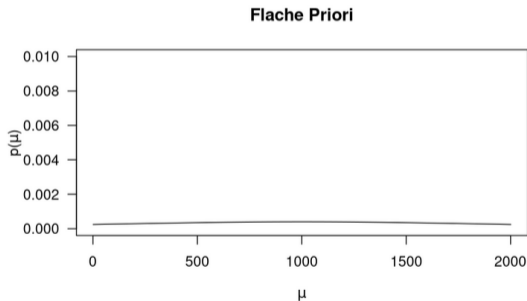
$$\lim_{\sigma_0^2 \rightarrow \infty} \exp \left( -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

- Dies ist eine **improper** Verteilung, da sie nicht integrierbar ist.
- Entspricht auch der Laplace-Priori, also Gleichverteilung über ganz  $\mathbb{R}$ .
- Jeffreys' Priori ist ein Grenzfall der konjugierten Normalverteilung.

# Visualisierung: Flache Priori

Eine **relativ flache Priori** ergibt sich z.B. für:

$$\sigma_0^2 = 1,000,000$$



Alternativ zur echten Improper-Priori kann man in der Praxis eine sehr große Varianz wählen.

Wir vermuten, dass die Schokohasen einer bestimmten Marke weniger als das angegebene Gewicht von 100 Gramm haben. Wir kaufen uns 10 Hasen und wiegen sie. Folgende Gewichte beobachten wir:

100.37 103.04 99.34 101.5 97.44 100.26 99.19 100.12 99.76 98.51

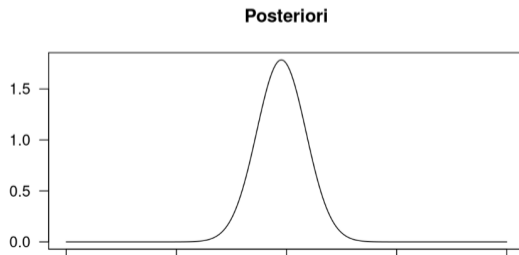
Im Folgenden wollen wir diese Daten Bayesianisch analysieren.

# Posteriori

Wählen wir im Folgenden also nun Jeffreys' Prior. Gegeben der Beobachtungen der Gewichte unserer Schokohasen (und der Herstellerangabe  $\sigma^2 = 0.5$ ) erhalten wir:

- $\tilde{m} = \frac{\sum_{i=1}^n x_i}{\sigma^2} = 1999.06$
- $\tilde{\tau}^2 = \frac{n}{\sigma^2} = \frac{10}{0.5} = 20$
- $\tilde{\mu} = \frac{\tilde{m}}{\tilde{\tau}} = 99.953$
- $\tilde{\sigma}^2 = \frac{1}{\tilde{\tau}} = 0.05$

Unsere Posteriori sieht also wie folgt aus:



Unsere Vermutung war, dass das Gewichts der Hasen im Mittel unter 100 Gramm liegt. Wir können nun die Posteriori-Wahrscheinlichkeit dafür angeben:

$$0.5832405$$

$$P(\mu \leq 100|x) \approx 0.5832405$$

Die Wahrscheinlichkeit dafür, dass der Erwartungswert kleiner als 100 Gramm ist, ist gegeben der Beobachtungen also 58.3%. Umgekehrt: die Wahrscheinlichkeit, dass unsere Annahme falsch ist, liegt bei 41.7%.

Die Wahrscheinlichkeit deutet zwar leicht darauf hin, dass unsere Annahme stimmt, aber ein Nachweis für unsere Annahme ist es nicht.

Wie ist es mit der Angabe des Herstellers. Er sagt: Das Erwartungswert des Gewichts liegt bei 101 Gramm. Darüber können wir direkt keine Aussage machen, weil die Wahrscheinlichkeit

$$P(\mu = 101|x) = 0.$$

$\mu|x$  ist eine stetige Zufallsvariable, dementsprechend hat jeder einzelne Wert die Wahrscheinlichkeit 0.

Alternativ können wir die Aussage “ $\mu$  ist 101 oder größer” überprüfen:

$$P(\mu \geq 101|x) = 1 - F(101) = 1.4182 \times 10^{-6}$$

# Unbekannte Varianz

In unserem Beispiel lässt sich nicht schlüssig nachweisen, dass der Erwartungswert niedriger als 100 ist, die Herstellerangabe scheint aber falsch. Könnte es aber sein, dass der Erwartungswert korrekt angegeben ist, die Varianz aber zu hoch ist?

## Normalverteilungsmodell

Wir nehmen wieder an:

$$X \sim N(\mu, \sigma^2).$$

Jetzt aber ist der Erwartungswert  $\mu = 101$  Gramm bekannt, die Varianz  $\sigma^2$  aber unbekannt. Welche Priori können wir für  $\sigma^2$  verwenden?

## Konjugierte Priori

Zuerst wollen wir wieder die konjugierte Priori finden. Der Kern der Normalverteilungsdichte für eine Beobachtung enthält nun folgende Teile (diesmal alle Teile, die von  $\sigma^2$  abhängen)

$$f(x_i | \sigma^2) \propto \frac{1}{\sigma^2} \exp\left(-\frac{1}{2\sigma^2} (x_i - \mu)^2\right) \propto (\sigma^2)^{-1} \exp\left(-(\sigma^2)^{-1} \frac{(x_i - \mu)^2}{2}\right)$$

Dieser Kern entspricht für  $\sigma^2$  der Form der Dichte einer Invers-Gamma-Verteilung:

$$p(\sigma^2) \propto (\sigma^2)^{a-1} \exp(-b/\sigma^2)$$

$\sigma^2 \sim IG(a, b)$  ist also die konjugierte Priori und führt zur Posteriori

$$\sigma^2 | x \sim IG\left(a + \frac{n}{2}, b + \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2\right)$$

# Konjugierte Priori für Präzision

Zur Vereinfachung der Notation ist es oft besser, statt der Varianz die inverse Varianz (genannt Präzision) zu betrachten:

$$\tau = \sigma^{-2}$$

Die konjugierte Priori von  $\tau$  ist dann die Gamma-Verteilung:

$$p(\tau) = \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau).$$

Die Posteriori lautet

$$p(\tau|x) \propto \tau^{n/2} \exp\left(-\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2\right) \tau^{a-1} \exp(-b\tau),$$

ist also die  $Ga(\frac{a+n}{2}, b + 0.5 \sum_{i=1}^n (x_i - \mu)^2)$ -Verteilung.

Als Jeffreys' Prior ergibt sich

$$p(\sigma^2) \propto \sigma^{-2}$$

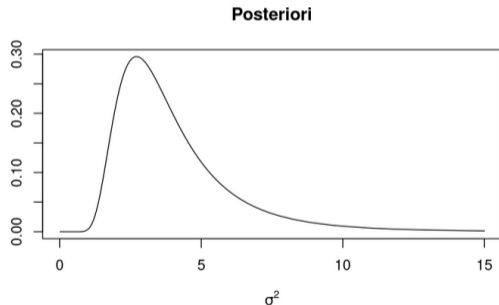
Erneut ist dies der Grenzfall der konjugierten Priori, nämlich für  $a \rightarrow 0$  und  $b \rightarrow 0$  (also “ $IG(0, 0)$ ”).

Im Gegensatz zu oben ist dies aber nicht eine Gleichverteilung auf  $\mathbb{R}^+$ .

# Posteriori-Wahrscheinlichkeit

Für unsere Beobachtungen ergibt sich  $\sum_{i=1}^n (x_i - \mu)^2 \approx 32.57$  (dabei ist  $\mu = 101$  bekannt) und mit der Jeffreys' Priori folgende Posteriori:

$$\sigma^2 | x \sim IG(10/2, 32.57/2)$$



Wir können also mit F die Verteilungsfunktion der  $IG(5, 16.28675)$ -Verteilung die Posteriori-Wahrscheinlichkeit für  $\sigma > 0.5$  berechnen.

## Normalverteilungsmodell mit zwei unbekannten Parametern

# Unabhängige konjugierte Prioris

Bei jeweils einem unbekannten Parameter und unter Benutzung der konjugierten Verteilung kennen wir die konjugierte Posteriori vollständig. Im Folgenden seien beide Parameter  $\mu$  und  $\sigma^2$  unbekannt. Zur einfacheren Notation benutzen wir wieder die Präzision  $\tau = 1/\sigma^2$ . Zur Erinnerung, konjugierte Prioris im eindimensionalen waren:

- $\mu \sim N(\mu_0, \sigma_0^2)$
- $\tau \sim Ga(a, b)$

## Gemeinsame Priori

Als ersten Ansatz folgende Idee: Wir wollen die konjugierten Prioris aus dem eindimensionalen benutzen und gehen außerdem “a priori” von Unabhängigkeit der Parameter aus. Die gemeinsame Priori-Dichte ist dann das Produkt der einzelnen

$$p(\mu, \tau) = p(\mu)p(\tau)$$

Die Posteriori lautet dann bis auf Konstanten:

$$p(\mu, \tau) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \quad (4)$$

$$\tau^{n/2} \exp\left(\tau/2 \sum_{i=1}^n (x_i - \mu)^2\right) \tau^{a-1} \exp(-b\tau) = \quad (5)$$

$$\tau^{n/2+a-1} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 - \tau/2 \sum_{i=1}^n (x_i - \mu)^2 - b\tau\right) \quad (6)$$

- Diese Posteriori hat nicht die Form der Priori. Die beiden Parameter  $\mu$  und  $\tau$  sind a posteriori voneinander abhängig, a priori aber nicht.
- Das Produkt der beiden konjugierten Priori-Dichten ergibt also keine konjugierte Priori! Vielmehr ist die entstandene Posteriori keine bekannte zweiparametrische Verteilung.
- Wir werden später zu diesem Ansatz zurückkommen, versuchen aber erstmal, eine konjugierte zweidimensionale Priori zu finden.

# Konjugierte zweidimensionale Priori

Ein anderer Ansatz beruht darauf, dass die Priori-Verteilung eines Parameters vom anderen Parameter abhängt. Genauer gesagt konstruieren wir eine bedingte Priori-Verteilung von  $\mu$  gegeben  $\tau$ :

$$\mu|\tau \sim N(\mu_0, 1/\tau)$$

Für  $\tau$  geben wir eine marginale Priori-Verteilung vor:

$$\tau \sim Ga(a, b)$$

- Wir übernehmen also die konjugierten Verteilungen (Normalverteilung bzw. Gammaverteilung), aber nehmen keine Unabhängigkeit der Parameter an.
- Die Begründung für diesen Ansatz ist wie folgt: Bei größerer Varianz (kleinerer Präzision) in den Daten brauchen wir auch größere Varianz (kleinere Präzision) in der Priori von  $\mu$ , da sonst die Priori zu informativ wird.

Die gemeinsame Priori  $p(\mu, \tau)$  ergibt sich nach der Definition der bedingten Dichte ebenfalls als Produkt der bedingten Dichte von  $\mu$  gegeben  $\tau$  und der marginalen Dichte  $\tau$ :

$$p(\mu, \tau) = p(\mu|\tau)p(\tau)$$

Die gemeinsame Priori-Dichte hat die Form

$$p(\mu, \tau) = \frac{b^a \sqrt{\lambda}}{\Gamma(a) \sqrt{2\pi}} \tau^{a-1/2} \exp(-b\tau) \exp(-\lambda$$

$$\tau (\mu - \mu_0)^2 \frac{1}{2}$$

Diese sogenannte Normal-Gamma-Verteilung mit Parametern  $(\mu_0, \lambda, a, b)$  ist in der Tat die konjugierte Priori zur zweiparametrischen Normalverteilung. Als Posteriori ergibt sich:

$$p(\mu, \tau | x) = \text{NormalGamma} \left( \frac{\lambda \mu_0 + n \bar{x}}{\lambda + n}, \lambda + n, a + \frac{n}{2}, b + \frac{1}{2} \left( ns^2 + \frac{\lambda n (\bar{x} - \mu_0)^2}{\lambda + n} \right) \right)$$

mit  $n$  der Anzahl an Beobachtungen,  $\bar{x} = \sum_{i=1}^n x_i / n$  dem Mittelwert der Beobachtungen und  $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$  der empirischen Varianz.

Die Priori-Parameter können wie folgt interpretiert werden:

- $\mu_0$  ist der Priori-Erwartungswert von  $\mu$
- $a/b$  ist der Priori- Erwartungswert von  $\tau$
- $a/b^2$  ist die Priori-Varianz von  $\tau$
- $\lambda$  ist die “Anzahl von Beobachtungen” in der Priori, vergleiche dazu das Binomialmodell

Für  $\lambda \rightarrow 0$  geht die bedingte Priori von  $\mu$  gegen die Gleichverteilung auf ganz  $\mathbb{R}$ .

Posteriori-Erwartungswert und Posteriori-Modus lassen sich direkt aus den Parametern ablesen.  
Der Posteriori-Erwartungswert ist

$$\mathbb{E}(\mu, \tau | x) = \left( \frac{\lambda \mu_0 + n \bar{x}}{\lambda + n}, \frac{a + \frac{n}{2}}{b + \frac{1}{2} \left( ns^2 + \frac{\lambda n (\bar{x} - \mu_0)^2}{\lambda + n} \right)} \right)$$

MAP

$$(\hat{\mu}, \hat{\tau})_{MAP} = \left( \frac{\lambda \mu_0 + n \bar{x}}{\lambda + n}, \frac{a + \frac{n-1}{2}}{b + \frac{1}{2} \left( ns^2 + \frac{\lambda n (\bar{x} - \mu_0)^2}{\lambda + n} \right)} \right)$$

- Kommen wir nochmal zurück zur Unabhängigkeitspriori, also  $p(\mu, \tau) = p(\mu)p(\tau)$ . Statt der gemeinsamen Posteriori-Verteilung sehen wir uns nun bedingte und marginale Posteriori an.
- Im vorliegenden Fall haben wir drei unterschiedliche Zufallsgrößen:  $\mu, \tau$  und die Daten  $x$ .
- Betrachten wir die bedingte Verteilung von  $\mu$  gegeben  $\tau$  und  $x$ , dann nennen wir dies **bedingte Posteriori** des Parameters  $\mu$ . Posteriori bezeichnet immer die Bedingung auf die Beobachtungen  $x$ , der Zusatz bedingt bezieht sich hier also darauf, dass wir bezüglich eines anderen Parameters, hier also  $\tau$  bedingen. Alle Verteilungen, die wir im Folgenden betrachten, sind immer gegeben der Daten  $x$ .

Nach der Definition der bedingten Dichte gilt

$$p(\mu|\tau, x) = p(\mu, \tau|x)p(\tau|x) \propto p(\mu, \tau|x)$$

Nach dem Satz von Bayes gilt weiterhin

$$p(\mu|\tau, x) \propto f(x|\mu, \tau)p(\mu, \tau)$$

und da  $\mu$  und  $\tau$  a priori unabhängig:

$$p(\mu|\tau, x) \propto f(x|\mu, \tau)p(\mu)p(\tau) \propto f(x|\mu, \tau)p(\mu)$$

Gegeben  $\tau$  ist also die Normalverteilungspriori quasi die konjugierte Priori zur bedingten Priori von  $\mu$ . Die Herleitung kennen wir aus dem eindimensionalen Fall, bei dem ja  $\tau$  gegeben war.

Wie bereits beschrieben ist die Unabhängigkeitspriori keine konjugierte Priori. Allerdings ist die Priori von  $\mu$  eine semikonjugierte Priori zur **bedingten Posteriori**.

## Definition:

Eine Familie  $\mathcal{F}$  von Verteilungen auf  $\Theta$  heißt semikonjugiert wenn für jede Priori  $p(\theta)$  auf  $\mathcal{F}$  die vollständig bedingte Posteriori  $p(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p; x)$  ebenfalls zu  $\mathcal{F}$  gehört. Hier ist die bedingte Posteriori-Dichte von  $\mu$  gegeben  $\tau$  also die Normalverteilung. Die bedingte Posteriori hilft uns aber erst mal nicht weiter, weil wir den Parameter  $\tau$  ja nicht kennen.

# Marginale Posteriori

Wir können nun für  $\tau$  die marginale Posteriori betrachten, also die Verteilung von  $\tau|x$ . Diese erhalten wir durch marginalisieren der gemeinsamen Posteriori

$$p(\tau|x) = \int p(\mu, \tau|x) d\mu.$$

Einfacher ist es, die Definition der bedingten Dichte zu benutzen und umzustellen:

$$p(\tau|x) = \frac{p(\mu, \tau|x)}{p(\mu|\tau, x)}$$

Durch einsetzen erhalten wir:

$$p(\tau|x) \propto \frac{f(x|\mu, \tau)p(\tau)}{p(\mu|\tau, x)}$$

Die einzelnen Terme sind bekannt

- $f(x|\mu, \tau)$ , die Datendichte einer Normalverteilung  $N(\mu, 1/\tau)$
- $p(\tau)$  die Prioridichte einer Gammaverteilung  $\text{Ga}(a, b)$
- $p(\mu|\tau; x)$  die Dichte einer Normalverteilung  $N(\tilde{\mu}, 1/\tilde{\tau})$  mit  $\tilde{\tau} = n/\sigma^2 + \sigma_0^{-2}$  und  $\tilde{\mu} = \tilde{\tau}(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \mu_0/\sigma_0^2)$  (siehe einparametrische Normalverteilung mit bekannter Varianz).

Die marginale Posteriori von  $\tau$  ist keine Standard-Verteilung. Wir können aber Zufallszahlen aus dieser Verteilung ziehen. Dieses Ziehen aus der Posteriori ist ein zentrales Konzept bei der hochdimensionalen Bayes-Statistik. Numerische Verfahren, die auf dem Ziehen von Zufallszahlen beruhen, werden allgemein Monte-Carlo-Verfahren bezeichnet, nach der Spielbank in Monte-Carlo im Fürstentum Monaco. Später werden wir Markov Chain Monte Carlo (MCMC)-Verfahren kennen lernen, eines der wichtigsten Hilfsmittel der Bayes-Statistik sind.

Theoretischer Hintergrund von Monte-Carlo-Verfahren ist das Gesetz der großen Zahlen. Dieser sagt (unter anderem) aus, dass für  $n \rightarrow \infty$  die empirische Verteilungsfunktion punktweise gegen die wahre Verteilungsfunktion geht. Ziehen wir also oft genug Zufallszahlen aus der Verteilung, können wir die wahre Verteilungsfunktion beliebig genau approximieren.

Hier können wir den CDF-Sampler benutzen. CDF steht für cumulative distribution function, also (kumulative) Verteilungsfunktion.

Idee dabei ist, den Träger der Verteilung zu diskretisieren.

Diskretisiere den Träger der zu simulierenden Verteilung in eine Menge von  $N$  Punkten

$$x_1 \leq \dots \leq x_N.$$

Evaluere die bis auf Proportionalität bekannte Dichte an  $x_1 \leq \dots \leq x_N$ , um Werte  $f_1, \dots, f_N$  zu erhalten.

Schätze die Proportionalitätskonstante  $c$  über  $c = f_1 + \dots + f_N$ . Ziehe Zufallszahlen aus  $x_1 \leq \dots \leq x_N$  gemäß den Wahrscheinlichkeiten  $f_1/c, \dots, f_N/c$ .

Mittels des CDF-Sampler, wie mit anderen Monte Carlo-Methoden, erhalten wir Ziehungen aus der Posteriori. Aus diesen können wir die Posteriori-Dichte der Parameter  $\mu$  und  $\tau$  schätzen: Punkt- und Intervallschätzer von einzelnen Parametern erhalten wir also immer aus den marginalen Posteriori-Verteilungen.

Ist nun der Mittelwert des Geschichts der Schokohasen kleiner als 101 Gramm. Und ist die Varianz größer als 0.5? Die Posteriori-Wahrscheinlichkeiten schätzen wir über die relativen Häufigkeiten dieser Fälle unter unseren Ziehungen. Dabei ist  $m$  die Anzahl der Ziehungen. Die Posteriori-Wahrscheinlichkeit, dass  $\mu < 101$  Gramm oder mehr ist, liegt also gerade mal bei 8.6%: eine geringe Wahrscheinlichkeit, aber eventuell ist das Ergebnis doch einfach auf die zufällige Auswahl der Schokohasen zurückzuführen.

Die Posteriori-Wahrscheinlichkeit, dass  $\sigma^2$  größer als 0.5 ist, ist dagegen 100% (in allen Ziehungen ist  $\sigma^2$  größer als 0.5). Wir sind uns also sehr sicher, dass die Varianz in den Gewichten zu hoch ist.

(Zur Transparenz: es handelt sich nicht um reale Messungen, die Gewichte wurden aus einer  $N(101, 2)$ -Verteilung simuliert.)

# Numerische Approximation

Alternativ lassen sich die marginalen Posterioris beider Parameter über Diskretisierung approximieren und man kann auf das Ziehen komplett verzichten.

Dabei kann man folgenden Weg gehen: Wir hatten oben  $p(\mu|x)$  hergeleitet. Nun gilt

$$p(\tau|x) = \int (\mu, \tau|x) d\mu = \int (\tau|\mu; x) p(\mu|x) d\mu$$

Für jedes feste  $\tau$  ist

$$E_{\mu|x}(p(\tau|\mu; x)) = \int p(\tau|\mu; x) p(\mu|x) d\mu = p(\tau|x)$$

Die marginale Posteriori-Dichte an einem Punkt  $\tau$  erhalten wir also als Erwartungswert der bedingten Dichte am Punkt  $\tau$ , wobei der Erwartungswert bezüglich der marginalen Posteriori von  $\mu|x$  gebildet wird.

Da wir  $\mu$  diskretisiert haben, können wir den Erwartungswert leicht berechnen (siehe oben). Dies machen wir auf einem Gitter für  $\tau$ :

Wir haben in diesem Abschnitt erstmals mehr als einen unbekannten Parameter betrachtet.

## **Priori**

Für die Konstruktion von Prioris haben wir zwei Wege benutzt:

- Unabhängigkeit der Parameter, gemeinsame Priori ist dann Produkt der einzelnen marginalen Prioris.
- Bedingte Priori eines Parameters gegeben einem anderen. Dann ergibt sich die gemeinsame Prior als Produkt von bedingter und marginaler Priori.
- Beide Ansätze lassen sich auf mehr als zwei Dimensionen verallgemeinern.
- Später werden wir noch Modelle sehen, bei denen wir Abhängigkeiten in die Priori-Verteilungen zulassen.

## Posteriori

Die Berechnung der Posteriori erfolgte im letzten Ansatz nicht mehr analytisch, sondern numerisch. Zwei Ansätze haben wir dabei erfolgt:

- Monte Carlo-Verfahren: Wir ziehen aus der gemeinsamen Posteriori-Verteilung, über das Gesetz der großen Zahlen erhalten wir dann Punktschätzer, Intervallschätzer und Posteriori-Wahrscheinlichkeiten.
- Numerische Approximation der Posteriori. Hier vor allem über Diskretisierung der marginale Posterioris.

Beide Ansätze werden in höherdimensionalen Modellen auch angewandt.

Die Schweiz trat Ende des 19. Jahrhunderts in eine Periode ein, die als demographischer Übergang bezeichnet wird, d.h. ihre Fruchtbarkeit begann von dem für unterentwickelte Länder typischen hohen Niveau abzufallen. Francine Vanderwalle sammelte Daten über die Fruchtbarkeit und verschiedene Kovariablen in der Schweiz in dieser Zeit. Wir benutzten die Daten aus dem Jahr 1888 für die 47 französischsprachigen Bezirke.

Wir wollen die Abhängigkeit des Fertilitätsindex vom Anteil des Agrarsektors (genauer: Anteil der Arbeitsplätze (von Männern) in der Landwirtschaft untersuchen. Wir gehen von einem linearen Zusammenhang aus und benutzten lineare Regression.

Das übliche lineare Regressionsmodell für  $y$  mit einer Kovariablen  $x$  ist

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

Dabei werden in der Regel folgende Annahmen an den Fehler  $\epsilon_i$  getroffen:

$$\mathcal{E}(\epsilon_i) = 0 \tag{7}$$

$$\text{Var}(\epsilon_i) = \sigma^2 \tag{8}$$

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \tag{9}$$

Zusätzlich nimmt man oft an

$$\epsilon_i \sim N(0, \sigma^2).$$

# Bayesianisches lineares Regressionsmodell

Für einen Bayesianischen Ansatz brauchen wir eine Datendichte, also eine Verteilungsannahme. Aus dem linearen Modell  $y_i = \alpha + \beta x_i + \epsilon_i$  und der Annahme  $\epsilon_i \sim N(0, \sigma^2)$  ergibt sich eine Verteilung für  $y$

$$y_i | \alpha, \beta, \sigma^2 \sim N(\alpha + \beta x_i, \sigma^2).$$

Genauer gesagt ist dies die bedingte Verteilung von  $y_i$  gegeben den Parametern  $\alpha, \beta$  und  $\sigma^2$ . In der Bayesianischen Statistik sind die Parameter ja Zufallsvariablen und haben eine Verteilung: Die Priori-Verteilung, die wir noch festlegen müssen, bzw. die Posteriori-Verteilung gegeben der Daten.

# Bayesianisches lineares Regressionsmodell

Die Datendichte aller  $y = (y_1, \dots, y_n)$  gegeben den Parametern  $\theta = (\alpha, \beta, \sigma^2)$  ist damit:

$$f(y|\theta) = \prod_{i=1}^n \frac{1}{(2\sigma^2)^{1/2}} \exp\left(-\frac{1}{2}\sigma^2(y_i - \alpha - \beta x_i)^2\right) \quad (10)$$

$$= \frac{1}{(2\sigma^2)^{n/2}} \exp\left(-\frac{1}{2}\sigma^2 \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \quad (11)$$

# Priori der Regressionsparameter

Bei der Wahl der Priori lehnen wir uns an der Wahl der Priori im Normalverteilungsmodell an. Dort hatten wir eine Normalverteilungspriori für den Erwartungswert benutzt. Nehmen wir an, dass

- $\alpha \sim N(m_\alpha, \nu_\alpha^2)$
- $\beta \sim N(m_\beta, \nu_\beta^2)$

und  $\alpha$  und  $\beta$  a priori unabhängig. Dann gilt

$$(\alpha + \beta x_i) \sim N(m_\alpha + x_i m_\beta, \nu_\alpha^2 + x_i^2 \nu_\beta^2)$$

- Wie wir später zeigen werden, handelt es sich hierbei um die semi-konjugierten Prioris für  $\alpha$  und  $\beta$ .
- Für eine möglichst wenig informative Priori setzten wir die Priori-Varianzen  $\nu_\alpha^2$  und  $\nu_\beta^2$  sehr hoch (flache Priori) oder lassen sie gar gegen  $+$  gehen. Dann haben wir uneigentliche Verteilungen:
- $p(\alpha) \propto \text{const.}$
- $p(\beta) \propto \text{const}$

Analog benutzen wir für die Varianz  $\sigma^2$  die (semi-)konjugierte Priori:

$$\sigma^2 IG(a, b),$$

oder, basierend auf der eindimensionalen Jeffreys' Priori

$$p(\sigma^2) \propto \sigma^{-2},$$

welche wieder eine uneigentliche Priori ist.

Zusätzlich gehen wir a priori davon aus, dass  $\sigma^2$  von  $\alpha$  und  $\beta$  stochastisch unabhängig ist. Damit erhalten wir die gemeinsame Priori-Dichte aller Parameter:

$$p(\alpha, \beta, \sigma^2) = p(\alpha)p(\beta)p(\sigma^2)$$

Die Annahme der Priori-Unabhängigkeit der Parameter ist Standard und wird implizit immer unterstellt, solange keine konkrete Abhängigkeit in die Parameter modelliert wird.

Leiten wir nun die Posteriori der Parameter gegeben der Daten her. Es gilt mit Daten  $y = (y_1, \dots, y_n)$  und Parametern  $\theta = (\alpha, \beta, \sigma^2)$ :

$$p(\theta|y) \propto f(y|\theta)p(\theta) \quad (12)$$

$$\propto \prod_{i=1}^n f(y_i|\theta)p(\alpha)p(\beta)p(\sigma^2) \quad (13)$$

$$\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \quad (14)$$

$$\nu_{\alpha}^{-1} \exp\left(-\frac{1}{2\nu_{\alpha}^2} (\alpha - m_{\alpha})^2\right) \quad (15)$$

$$\nu_{\beta}^{-1} \exp\left(-\frac{1}{2\nu_{\beta}^2} (\alpha - m_{\beta})^2\right) \sigma^{-2(\alpha+1)} \exp(b\sigma^{-2}) \quad (16)$$

Diese Posterioridichte ist wohl keine Standard-Verteilung mehr. Wie können wir die gemeinsame Posteriori erhalten, wie die marginalen Posterioris, aus denen wir Schätzer ableiten?

Im Folgenden benutzen wir den sogenannten Gibbs-Sampler. Hierbei handelt es sich wieder um einen Monte-Carlo-Algorithmus. Ziel ist es, Zufallszahlen aus der Posterioriverteilung zu ziehen und:

- den Posteriori-Erwartungswert über den Mittelwert der gezogenen Zufallszahlen schätzen;
- die Posteriori-Wahrscheinlichkeit eines Intervalls über die relative Häufigkeit zu schätzen, mit der die gezogenen Zufallszahlen in das Intervall fallen;
- die Posteriori-Dichte über ein Histogramm oder mittels des Kern-Dichte-Schätzers zu schätzen.

## Idee

Die Grundidee des Gibbs-Samplers ist es, abwechselnd aus der bedingten (Posteriori-)Verteilung eines Parameters gegeben der anderen Parameter zu ziehen.

Gegeben sei die Dichte  $f(\theta)$  mit  $\theta = (\theta_1, \dots, \theta_p)$ . Der Algorithmus des Gibbs-Samplers ist wie folgt:

- 1 Lege Startwerte  $\theta_i^{(0)}$  für alle  $i = 1, \dots, p$  fest.
- 2 Setze Iteration  $k = 1$ .
- 3 Ziehe einen Wert  $\theta_1^{(k)}$  gegeben  $\theta_2^{(k-1)}, \dots, \theta_p^{(k-1)}$  aus der Dichte  $f(\theta_1 | \theta_2, \dots, \theta_p)$ .
- 4 Ziehe analog für  $j = 2, \dots, p$  einen Wert  $\theta_j^{(k)}$  gegeben  $\theta_1^{(k)}, \dots, \theta_{j-1}^{(k)}, \theta_{j+1}^{(k-1)}, \dots, \theta_p^{(k-1)}$  aus der Dichte  $f(\theta_j | \theta_{-j})$ .
- 5 Erhöhe  $k$  um 1.
- 6 Wiederhole Schritte 3 bis 5 so oft wie nötig.

Dabei ist  $(\theta_{-j} = \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p)$ , also alle Parameter außer  $\theta_j$ .  $\theta_j^{(k)}$  steht für den Wert, den  $\theta_j$  in der Iteration  $k$  hat.

- Die Ziehungen sind beim Gibbs-Sampler voneinander abhängig:  $\theta_1^{(k+1)}$  hängt von  $\theta_{-1}^{(k)}$  ab, die wiederum von  $\theta_1^{(k)}$  abhängen. Damit hängt  $\theta_1^{(k+1)}$  also vom vorherigen Wert  $\theta_1^{(k)}$  ab.
- Trotz der Abhängigkeit der Ziehungen kann man zeigen, dass der Algorithmus Zufallszahlen aus der gemeinsamen Dichte  $f(\theta)$  zieht.
- Die Schätzung ist allerdings weniger effizient, als wenn wir unabhängige Ziehungen hätten. Dass heißt, wir müssen insgesamt öfter ziehen.
- Zu Beginn des Algorithmus hängen die Ziehungen zudem von den im Schritt 1 gewählten Startwerten ab. Wir müssen die Ziehungen am Anfang also ignorieren – die sogenannte **burn-in**-Phase.
- Nach dem burn-in können wir Erwartungswert, Varianz, Wahrscheinlichkeiten, Dichte etc. schätzen (trotz der Abhängigkeit der Ziehungen).
- Dafür braucht man eine allgemeinere Version des Gesetzes der großen Zahlen, den Ergodensatz.

# Vollständig bedingte Posterioris

- Kommen wir zur Anwendung des Gibbs-Samplers in unserem Regressionsmodell.
- Wir brauchen für den Algorithmus die Verteilung - in unserem Fall die Posteriori-Verteilung - eines Parameter ( $\theta_i$ ) gegeben den restlichen.
- Diese Verteilung nennen wir **vollständig bedingte Posteriori** oder englisch **full conditional**.

Wie können wir die full conditionals (man benutzt den Begriff sowohl für die Verteilung als auch für die Dichte) herleiten? Aus der Definition der bedingten Dichte ergibt sich:

$$f(\theta_i|\theta_{-i}) = f(\theta)/f(\theta_{-i}) \propto f(\theta)$$

Die Proportionalität ergibt sich dabei, da  $f(\theta_{-i})$  nicht von  $\theta_i$  abhängt, also bezüglich  $\theta_i$  eine Konstante ist.

Die vollständige bedingte Dichte ist also einfach proportional zur gemeinsamen Dichte.

- Praktisch nehmen wir uns aus der Posterioridichte einfach die Terme, die von  $\theta_i$  abhängen (und hoffentlich kommt dann eine Verteilung heraus, aus der wir ziehen können).
- Um aus  $f(\theta_i|\theta_{-i})$  zu ziehen, können wir entweder:
  - aus  $f(\theta)$  auf den Kern einer (Standard-)Verteilung schließen
  - oder ein approximatives Verfahren wie den CDF-Sampler benutzen (ist in der Regel aber langsamer)

# Full conditional von $\beta$

Die Posterioridichte von  $\theta = (\alpha, \beta, \sigma^2)$  ist

$$p(\theta|y) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \quad (17)$$

$$\nu_\alpha^{-1} \exp\left(-\frac{1}{2\nu_\alpha^2} (\alpha - m_\alpha)^2\right) \nu_\beta^2 \exp\left(-\frac{1}{2\nu_\beta^2} (\beta - m_\beta)^2\right) \quad (18)$$

$$\sigma^{-2(a+1)} \exp(b\sigma^{-2}) \quad (19)$$

Nehmen wir nun die Terme, die von  $\beta$  abhängen:

$$p(\beta|\alpha, \sigma^2; y) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \exp\left(-\frac{1}{2\nu_\beta^2} \beta(\beta - m_\beta)^2\right) \quad (20)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2} \left(\beta^2 \sum_{i=1}^n x_i^2 - 2\beta \sum_{i=1}^n (x_i(y_i - \alpha))\right) - \frac{1}{2\nu_\beta^2} \beta(\beta^2 - 2\beta m_\beta)\right) \quad (21)$$

$$\propto \exp\left(-\frac{1}{2} \beta^2 (\nu_\beta^{-2} + \sigma^{-2} \sum_{i=1}^n x_i^2) + \beta (m_\beta \nu_\beta^2 + \sigma^{-2} \sum_{i=1}^n (x_i(y_i - \alpha)))\right) \quad (22)$$

Durch quadratische Ergänzung erhält man:

- $\beta|\alpha, \sigma^2; y$  ist normalverteilt mit
- Inverser Varianz  $\tilde{\nu}_\beta^{-2} = \nu_\beta^{-2} + \sigma^{-2} \sum_{i=1}^n x_i^2$
- und Erwartungswert  $m_\beta = \tilde{\nu}_\beta^{-2}(\nu_\beta^{-2} m_\beta + \sigma^{-2} \sum_{i=1}^n (x_i(y_i - \alpha)))$

Analog erhält man

$$p(\alpha|\beta, \sigma^2; y) \propto \exp\left(-\frac{1}{2}\alpha^2(\nu_\alpha^{-2} + n\sigma^{-2}) + \alpha(m_\alpha + \sum_{i=1}^n(y_i - \beta x_i))\right)$$

Die Full Conditional ist also eine Normalverteilung mit

- Inverser Varianz  $\tilde{\nu}_\alpha^{-2} = \nu_\alpha^{-2} + n\sigma^{-2}$
- und Erwartungswert  $\tilde{m}_\alpha = \tilde{\nu}_\alpha^{-2}m_\alpha + \sigma^{-2}\sum_{i=1}^n(y_i - \beta x_i)$ .

# Full conditional von $\sigma^2$

Zuletzt leiten wir noch die Full Conditional von  $\sigma^2$  her:

Die Posterioridichte ist

$$p(\theta|y) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \nu_{\alpha}^{-1} \exp\left(-\frac{1}{2\nu_{\alpha}^2} (\alpha - m_{\alpha})^2\right) \quad (23)$$

$$\nu_{\beta}^{-1} \exp\left(-\frac{1}{2\nu_{\beta}^2} (\beta - m_{\beta})^2\right) \sigma^{-2} (a+1) \exp(b\sigma^{-2}) \quad (24)$$

Folgende Terme hängen von  $\sigma^2$  ab:

$$p(\sigma^2|\alpha, \beta; y) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \sigma^{-2} (a+1) \exp(b\sigma^{-2}) \quad (25)$$

$$\propto \sigma^{2(a+n/2-1)} \exp\left(\sigma^{-2} \left(b + \frac{1}{2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right)\right) \quad (26)$$

Das ist der Kern einer Invers-Gammaverteilung mit Parametern

- $\tilde{a} = a + n/2$
- $\tilde{b} = b + 0.5 \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$

Wir erinnern uns, dass wir (angeblich) semi-konjugierte Prioris benutzt haben. Semi-Konjugiertheit hatten wir wie folgt definiert:

*Die vollständig bedingte Posteriori eines Parameters hat die selbe Verteilung wie die Priori des Parameters, bis auf andere Parameter.*

Die Prioris sind also in der Tat normalverteilt, die full conditionals entsprechen den Priori Verteilungen ist es nicht verwunderlich, dass wir auf diese full conditionals gekommen sind:

- $\alpha \rightarrow$  Normalverteilung
- $\beta \rightarrow$  Normalverteilung
- $\sigma^2 \rightarrow$  Invers-Gamma-Verteilung

Sehen wir uns die Ergebnisse der Ziehungen an. Die folgenden Graphiken zeigen die Ziehungen der drei Parameter über 20 Iteration, die sogenannten Trace Plots: Burn-In

Wir sehen, dass die Zufallszahlen in den ersten paar Iterationen einen sehr starken Trend aufweisen. Danach sind die Ziehungen annähernd stabil aus einem bestimmten Bereich. Diesen Bereich am Anfang nennen wir den burn-in. Der burn-in hängt von verschiedenen Eigenschaften des Algorithmus ab, insbesondere aber von den Startwerten. Theoretisch können die Startwerte beliebig gewählt werden, für ein große Anzahl von Iterationen erhalten wir Ziehungen aus der richtigen Verteilung. Ein “falscher” Startwert kann aber dazu führen, dass der burn-in länger wird. Im Beispiel wurde  $\beta_0 = -10$  gesetzt, so dass der Algorithmus einige Iterationen braucht, um in den “richtigen Bereich” zu kommen. Die korrekte Bestimmung des burn-in, also ab wann der Algorithmus wirklich Ziehungen aus der gemeinsamen Verteilung liefert, ist nicht trivial. Wir werden später sogenannten Kovergenzdiagnostiken besprechen. Vorerst bestimmen wir denn burn-in visuell und großzügig. Hier legen wir ihn auf 100 Iterationen fest, sprich wir löschen die ersten 100 Iterationen:

Eine andere Frage ist, wie viele Iterationen wir brauchen. Die Antwort auf diese Frage hängt stark davon ab, welche Schlüsse man ziehen will.

Den Posteriori-Erwartungswert kann man über den Mittelwert der Ziehungen schätzen. Eine Möglichkeit ist, sich den Mittelwert über die bisherige Iterationen im Verlauf der Iterationen zu visualisieren (ohne den burn-in):

Wir sehen, dass der Mittelwert “konvergiert”, also mit steigender Anzahl von Iterationen stabil auf einen Wert zustrebt. Nach 900 Iterationen (ohne den burn-in) ändern sich (bei  $\beta$ ) die ersten drei Nachkommastellen des Mittelwerts nicht mehr. Die Anzahl der Iterationen ist hier (für den Posteriori-Erwartungswert) groß genug.

Aus den Ziehungen können wir nun Punkt- und Intervallschätzer berechnen. Konkret schätzen wir z.B. den Posteriori-Erwartungswert über den Mittelwert der Ziehungen und das symmetrische 95%-Kredibilitätsintervall über die 2.5%- und 97.5%-Quantile der Ziehungen:

- $\alpha$ : Posteriori-Erwartungswert: 59.437, Kredibilitätsintervall: (51.129, 68.654)
- $\beta$ : Posteriori-Erwartungswert: 0.208, Kredibilitätsintervall: (0.048, 0.359)
- $\sigma^2$ : Posteriori-Erwartungswert: 146.915, Kredibilitätsintervall: (96.358, 221.647)

Die Parameter lassen sich wie im linearen Regressionsmodell üblich interpretieren:  $\beta$  ist der lineare Einfluß der Kovariable “Anteil des Agrarsektors”. Pro Punkt, den der Anteil der Agrarsektors höher ist, ist der Fertilitätsindex im Mittel um 0.208 Punkte höher. Der Intercept liegt bei 59.437, also bei einem Anteil des Agrarsektors von 0 liegt der Fertilitätsindex im Mittel bei 59.437.

Das 95%-Kreditintervall von  $\beta$  umfasst nicht die 0, das heißt der Anteil des Agrarsektors hat “signifikanten” Einfluß auf den Fertilitätsindex. Für diese Frage kann man auch  $P(\beta \leq 0|y)$  durch die relative Häufigkeit von  $\beta \leq 0$  schätzen; diese liegt bei 0.0066667. Dies kann man als Bayesianische Entsprechung des p-Wertes interpretieren. Der Punktschätzer der Fehlervarianz liegt bei 146.915, bzw. des Standardfehlers bei 0.4560702. Diese sagt aus, wie stark die Beobachtungen vom Modell abweichen.

Schließlich können wir uns auch die marginalen Posterioris der drei Parameter ansehen. Dazu berechnen wir einen Kerndichteschätzer aus den Ziehungen. (Siehe Online Kurs)  
Bei der Posteriori von  $\beta$  sehen wir nochmal, dass die 0 (sprich: kein Einfluß der Kovariable) “am Rand” der Posteriori-Dichte liegt.