

Chapter 5

Bayes'sche Ansätze

5.1 Einführung

In der Statistik gibt es zwei verschiedene Lehrmeinungen. Es sind dies

1. die **Frequentisten** und
2. die **Bayesianer**.

Alle bisher ¹ vorgestellten statistischen Konzepte, so zum Beispiel **Least Squares**, **Maximum Likelihood**, **REML** und **BLUP** stammen aus dem Lager der Frequentisten.

Die Unterschiede zwischen Frequentisten und Bayesianern bestehen hauptsächlich in

- deren Verständnis von Wahrscheinlichkeiten
- deren Unterteilung von Modell- und Datenkomponenten
- deren Techniken zur Schätzung von Parametern

Die folgende Tabelle gibt eine Übersicht über die Unterschiede.

Was	Frequentisten	Bayesianer
Wahrscheinlichkeiten	Eigenschaften von Zufallsvariablen, welche bei grossen Stichproben eintreten	Mass für Informationsgehalt unabhängig von Stichprobengrösse
Modell- und Datenkomponenten	Unterscheidung zwischen Modellparametern und Daten. Parameter sind unbekannt, Daten sind bekannt. Fehlende Daten werden ignoriert	Unterscheidung zwischen unbekanntem und bekannten Grössen, unabhängig ob Parameter oder Daten. Fehlende Daten können simuliert werden
Schätzungen von Parametern	ML oder REML werden für Parameterschätzung verwendet	MCMC Zufallszahlen zur Approximation der gewünschten a posteriori Verteilungen

5.2 Das Lineare Modell

Die Bayes'sche Art der Parameterschätzung soll an einem einfachen linearen Modell gezeigt werden. Angenommen, wir betrachten das Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \tag{5.1}$$

¹Hier ist nicht nur diese Vorlesung sondern auch die Züchtungslehre und die angewandte Zuchtwertschätzung gemeint

wobei y_i die i -te Beobachtung einer Zielgrösse ist, β_0 für den Achsenabschnitt steht, x_1 eine erklärende Variable ist und ϵ_i für den Restterm steht. Für den Restterm nehmen wir an, dass deren Varianz konstant gleich σ^2 ist.

5.2.1 Bekannte und Unbekannte

Unter der Annahme, dass wir für die Zielgrösse y_i und die erklärende Variable x_1 keine fehlenden Daten haben, dann machen wir als Bayesianer folgende Einteilung in bekannte und unbekannte Grössen.

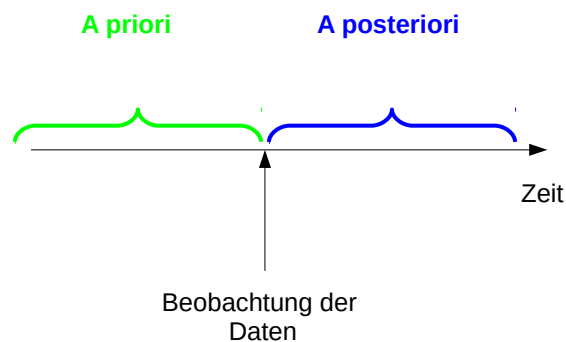
und als **bekannte** Grössen

Was	bekannt	unbekannt
y_i	X	
x_1	X	
β_0		X
β_1		X
σ^2		X

5.2.2 Vorgehen bei Parameterschätzung

Bayesianer basieren Schätzungen von unbekanntem Grössen auf der sogenannten **a posteriori Verteilung** der unbekanntem Grössen gegeben die bekannten Grössen. Die a posteriori Verteilung wird mithilfe des **Satzes von Bayes** aufgrund der a priori Verteilung der unbekanntem und aufgrund der Likelihood berechnet.

Die Bezeichnungen "a priori" und "a posteriori" beziehen sich immer auf den Zeitpunkt der Beobachtung der analysierten Daten. Die jeweiligen Verteilungen quantifizieren den Informationsstand zu den Unbekanntem um jeweiligen Zeitpunkt. Dieses Konzept soll anhand der folgenden Grafik verdeutlicht werden.



Für unser Beispiel des einfachen linearen Modells, definieren wir zuerst den Vektor β als

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Die Beobachtungen y_i fassen wir ebenfalls zu einem Vektor y zusammen. Für den Moment nehmen wir an, dass σ^2 bekannt sei. Eine Bayes'sche Parameterschätzung für β basiert dann auf der a posteriori Verteilung $f(\beta|y, \sigma^2)$ der Unbekanntem β gegeben die Bekannten y und σ^2 . Diese a posteriori Verteilung lässt sich mit dem Satz von Bayes, wie folgt berechnen

$$\begin{aligned}
f(\beta|y, \sigma^2) &= \frac{f(\beta, \sigma^2, y)}{f(y, \sigma^2)} \\
&= \frac{f(y|\beta, \sigma^2)f(\beta)f(\sigma^2)}{f(y, \sigma^2)} \\
&\propto f(y|\beta, \sigma^2)f(\beta)f(\sigma^2)
\end{aligned} \tag{5.2}$$

In Gleichung (5.2) konnten wir die a posteriori Verteilung $f(\beta|y, \sigma^2)$ als Produkt der a priori Verteilungen ($f(\beta)$ und $f(\sigma^2)$) der unbekannt Grössen β und σ^2 und der Likelihood $f(y|\beta, \sigma^2)$ ausdrücken. Der Faktor $f(y, \sigma^2)^{-1}$ (Term im Nenner) entspricht der sogenannten Normalisierungskonstanten und ist nicht weiter von Interesse. Somit wird die a posteriori Verteilung oft als Proportionalitätsbeziehung angegeben.

Die a posteriori Verteilung $f(\beta|y, \sigma^2)$ ist in vielen Fällen nicht explizit darstellbar. Das war lange ein Problem, welches die Anwendung von Bayes'schen Analysen sehr einschränkte. Zwei Entwicklungen haben dieses Problem beseitigt.

1. In seinem Paper (Besag (1974)) zeigte Julian Besag, dass jede posteriori Verteilung durch eine Serie von Zufallszahlen aus den voll-bedingten Verteilungen bestimmt ist. Für unser Beispiel lauten die voll-bedingten Verteilungen: Bedingte Verteilung von β_0 gegeben alle anderen Grössen: $f(\beta_0|\beta_1, \sigma^2, y)$, bedingte Verteilung von β_1 gegeben alle anderen Grössen: $f(\beta_1|\beta_0, \sigma^2, y)$.
2. Die Entwicklung von effizienten Pseudo-Zufallszahlen-Generatoren auf dem Computer.

5.3 Gibbs Sampler

Die Umsetzung der beiden oben aufgelisteten Punkte führt zu einer Prozedur, welche als **Gibbs Sampler** bezeichnet wird. Wenden wir den Gibbs Sampler auf einfaches lineares Regressionsmodell an, dann resultiert das folgende Vorgehen bei der Analyse. Unabhängig vom verwendeten Modell läuft die Konstruktion einer Gibbs Sampling Prozedur immer in den folgenden Schritten ab. Diese Schritte können für die meisten Analysen wie ein Kochbuchrezept verwendet werden.

1. Bestimmung der a priori Verteilungen für die unbekannt Grössen.
2. Bestimmung der Likelihood
3. Bestimmung der voll-bedingten Verteilungen

5.3.1 A priori Verteilungen

In unserem Beispiel handelt es sich dabei um $f(\beta)$ und $f(\sigma^2)$. In den meisten Fällen, wenn man das erste Mal eine bestimmte Art von Daten analysieren soll, empfiehlt es sich eine sogenannte uninformative a priori Verteilung zu wählen. Eine uninformative a priori Verteilung bedeutet einfach, dass deren Dichtewert überall gleich, also eine Konstante ist. Wenden wir zum Beispiel für die Unbekannte β eine uninformative a priori Verteilung an, dann bedeutet das, dass wir $f(\beta) = c$.

Alternativ zu der unformativen a priori Verteilung gibt es auch a priori Verteilungen für bestimmte unbekannt Grössen, welche als de-facto Standard akzeptiert sind. Ein Beispiel dafür ist die a priori Verteilung der unbekannt Restvarianz, welche üblicherweise als Inverse-Chi-Quadrat Verteilung angenommen wird.

5.3.2 Likelihood

Die Likelihood ist wie bei den Frequentisten als bedingte Verteilung ($f(y|\beta, \sigma^2)$) der Daten y gegeben die Parameter (β und σ^2). Falls keine Daten fehlen, dann ist die Bayes'sche Likelihood und die frequentistische Likelihood gleich.

5.3.3 Vollbedingte Verteilungen

Mit vollbedingten Verteilungen ist gemeint, dass für jede unbekannte Grösse die bedingte Verteilung gegeben alle anderen Grössen bestimmt wird. In unserem Beispiel des linearen Regressionsmodells haben wir zwei unbekannte Grössen β_0 und β_1 . Somit haben wir auch zwei vollbedingte Verteilungen. Unter der Annahme, dass unsere Daten (y) einer Normalverteilung folgen, resultieren die folgenden vollbedingten Verteilungen.

unbekannte Grösse	unbekannte Grösse	resultierende Verteilung
β_0	$f(\beta_0 \beta_1, \sigma^2, y)$	$\mathcal{N}\left(\hat{\beta}_0, \text{var}(\hat{\beta}_0)\right)$
β_1	$f(\beta_1 \beta_0, \sigma^2, y)$	$\mathcal{N}\left(\hat{\beta}_1, \text{var}(\hat{\beta}_1)\right)$

Aufgrund von Berechnungen, welche hier nicht gezeigt sind, können wir die oben aufgelisteten vollbedingten Verteilungen bestimmen. Die entsprechenden Verteilungen sind in der Kolonnen ganz rechts, welche mit "resultierende Verteilung" überschrieben ist, aufgelistet. Dabei steht $\mathcal{N}()$ für die Normalverteilung. Für die Erwartungswerte und Varianzen wird das Modell in Gleichung (5.1) leicht umformuliert.

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{x}\beta_1 + \epsilon \quad (5.3)$$

Aus dem obigen Modell bilden wir ein neues Modell, welches auf der rechten Seite der Gleichung nur von β_0 und ϵ abhängt. Da wir wissen, dass die Verteilung der Least Squares Schätzer eine Normalverteilung ist, werden wir diese für die Bestimmung der vollbedingten Verteilungen verwenden.

$$\mathbf{w}_0 = \mathbf{1}\beta_0 + \epsilon \quad (5.4)$$

wobei $\mathbf{w}_0 = \mathbf{y} - \mathbf{x}\beta_1$. Aufgrund des Modells in Gleichung (5.4) können wir den Least Squares Schätzer für β_0 aufstellen. Dieser lautet:

$$\hat{\beta}_0 = (\mathbf{1}^T \mathbf{1})^{-1} \mathbf{1}^T \mathbf{w}_0 \quad (5.5)$$

Die Varianz des Least Squares Schätzers für β_0 lautet:

$$\text{var}(\hat{\beta}_0) = (\mathbf{1}^T \mathbf{1})^{-1} \sigma^2 \quad (5.6)$$

Analog zu β_0 berechnen wir den Least Squares Schätzer für β_1 und dessen Varianz.

$$\hat{\beta}_1 = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{w}_1 \quad (5.7)$$

wobei $\mathbf{w}_1 = \mathbf{y} - \mathbf{1}\beta_0$

$$\text{var}(\hat{\beta}_1) = (\mathbf{x}^T \mathbf{x})^{-1} \sigma^2 \quad (5.8)$$

5.3.4 Umsetzung des Gibbs Samplers

Der Gibbs Sampler wird durch wiederholtes ziehen von Zufallszahlen aus den oben angegebenen vollbedingten Verteilungen umgesetzt. Das heisst, wir setzen für alle unbekanntes Grössen sinnvolle Startwerte ein. Für β_0 und β_1 wählen wir 0 als Startwert. Dann berechnen wir den Erwartungswert und die Varianz für die vollbedingte Verteilung von β_0 . Aus dieser Verteilung ziehen wir einen neuen Wert für β_0 . In einem zweiten Schritt berechnen wir den Erwartungswert und die Varianz für die vollbedingte Verteilung von β_1 , wobei wir für β_0 schon den neuen Wert einsetzen. Aus der Verteilung für β_1 ziehen wir einen neuen Wert für β_1 . Danach beginnen wir die Schritte wieder bei β_0 . Diese Schrittabfolge wiederholen wir 10000 mal und

speichern alle gezogenen Werte für β_0 und β_1 . Die Bayes'schen Parameterschätzungen entsprechen dann den Mittelwerten der gespeicherten Werte.

Der folgende R-Codeblock soll die Umsetzung des Gibbs Samplers für β_0 und β_1 als Programm zeigen. Der Einfachheit halber wurde σ^2 konstant $\sigma^2 = 1$ angenommen.

```
# ### Startwerte für beta0 und beta1
beta <- c(0, 0)
# ### Bestimmung der Anzahl Iterationen
niter <- 10000
# ### Initialisierung des Vektors mit Resultaten
meanBeta <- c(0, 0)
for (iter in 1:niter) {
  # Ziehung des Wertes des Achsenabschnitts beta0
  w <- y - X[, 2] * beta[2]
  x <- X[, 1]
  xpxi <- 1/(t(x) %*% x)
  betaHat <- t(x) %*% w * xpxi
  # ### neue Zufallszahl fuer beta0
  beta[1] <- rnorm(1, betaHat, sqrt(xpxi))
  # Ziehung der Steigung beta1
  w <- y - X[, 1] * beta[1]
  x <- X[, 2]
  xpxi <- 1/(t(x) %*% x)
  betaHat <- t(x) %*% w * xpxi
  # ### neue Zufallszahl fuer beta1
  beta[2] <- rnorm(1, betaHat, sqrt(xpxi))
  meanBeta <- meanBeta + beta
}
# ### Ausgabe der Ergebnisse
cat(sprintf("Achsenabschnitt = %6.3f \n", meanBeta[1]/iter))
cat(sprintf("Steigung = %6.3f \n", meanBeta[2]/iter))
```


Abkürzungen

Abbreviation	Meaning
QTL	Quatitative Trait Loci
GWAS	Genome Wide Association Study
GLMM	Generalized Linear Mixed Models
i.i.d.	independent, identically distributed
QQ	Quantil Quantil
LASSO	Least Absolute Shrinkage And Selection Operator
REML	Restricted oder Residual Maximum Likelihood
ML	Maximum Likelihood
MCMC	Markov Chain Monte Carlo

Bibliography

- Balding, D. J., Bishop, M., and Cannings, C., editors (2009). *Handbook of Statistical Genetics*. Wiley.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, (36):192–236.
- Bühlmann, P. and Mächler, M. (2014). Computational statistics. Vorlesungsunterlagen ETHZ, 2014.
- Garrick, D., Taylor, J., and Fernando, R. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, (41(1)):55.
- Hofer, A. (1990). *Schätzung von Zuchtwerten feldgeprüfter Schweine mit einem Mehrmerkmals-Tiermodell*. PhD thesis, ETH Zürich.
- James, G., Witten, D., Hastie, T., and Tibshirani, R., editors (2013). *An Introduction to Statistical Learning*. Springer.
- Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, (157):1819–1829.
- von Rohr, P. (2016). Züchtungslehre. Vorlesungsunterlagen ETHZ, HS2016.