ASMNW - Übung 6

Peter von Rohr

2017-04-27

Aufgabe 1: Kontrollfragen LASSO

Kontrollfrage 1:

a. Wieso brauchen wir Alternativen zu Least Squares?
b. Wie sehen die Alternativen zu Least Squares aus?

Kontrollfrage 2: Gegeben das einfache lineare Modell

$$y = Xb + e \tag{1}$$

a. Welche Anforderung bezüglich des Ranges der Matrix $\mathbf{X}$ besteht, damit Least Squares verwendet werden kann?
b. Aus welchem Grund besteht diese Anforderung aus 2a?

Aufgabe 2: Kontrollfragen Bayes

Kontrollfrage 1: Gegeben sei das folgende einfache lineare Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \tag{2}$$

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. Teilen Sie die Komponenten im Modell (2) in der folgenden Tabelle in bekannte und unbekannte Grössen ein.

Was	bekannt	unbekannt
y_i		
x_1		
β_0		
β_1		
σ^2		

Kontrollfrage 2:

Unter der Annahme, dass bei der Zielgrösse und der erklärenden Variablen keine Daten fehlen, welcher Einteilung bei den Frequentisten entspricht dann die Bayes'sche Einteilung in bekannte und unbekannte Grössen?

Aufgabe 3: Vergleich zwischen Bayes und Least Squares

Gegeben ist das Modell

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i \tag{3}$$

wobei

y_i	Zielgrösse der Beobachtung i
β_0	Achsenabschnitt
x_i	erklärende Variable der Beobachtung i
β_1	fixer Parameter der erklärenden Variable x
ϵ_i	zufälliger Resteffekt der Beobachtung i

Wir nehmen an die Resteffekte seien unkorreliert und normalverteilt mit Erwartungswert 0 und konstanter Varianz $var(e_i) = \sigma^2$

Auf der Webseite ist unter dem Link https://charlotte-ngs.github.io/GELASMFS2017/w6/simpleLinReg.csv ein Datensatz mit 20 Beobachtungen verfügbar. Diesen Datensatz können Sie mit dem folgenden R-Befehl einlesen. Als Kontrolle können wir die Dimension der eingelesenen Daten bestimmen.

```
sDataFn <- "simpleLinReg.csv"
sDataLink <- file.path("https://charlotte-ngs.github.io/GELASMFS2017/w6", sDataFn)
dfDataRead <- read.csv(file = sDataLink, stringsAsFactors = FALSE)
dim(dfDataRead)
```

Mit dem folgenden Programm schätzen wir den Achsenabschnitt und den Koeffizienten der erklärenden Variablen mit einer Bayesschen Methode, welche auch als Gibbs Sampler bezeichnet wird.

```
### # Matrix X als Inzidenzmatrix des Achsenabschnitts und
### #  der erklärenden Variablen
X <- cbind(1,dfDataRead$x)
### # y als Vektor der Beobachtungen
y <- dfDataRead$y
### # Zuweisung der Startwerte
beta = c(0, 0)
# loop for Gibbs sampler
niter = 100000 # number of samples
meanBeta = c(0, 0)
for (iter in 1:niter) {
  # sampling intercept
  w = y - X[, 2] * beta[2]
  x = X[, 1]
  xpxi = 1/(t(x) %*% x)
  betaHat = t(x) %*% w * xpxi
  beta[1] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
  # sampling slope
  w = y - X[, 1] * beta[1]
  x = X[, 2]
  xpxi = 1/(t(x) %*% x)
  betaHat = t(x) %*% w * xpxi
  beta[2] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
  meanBeta = meanBeta + beta
  if ((iter%%20000) == 0) {
    cat(sprintf("Intercept = %6.3f \n", meanBeta[1]/iter))
    cat(sprintf("Slope = %6.3f \n", meanBeta[2]/iter))
```

```
  }
}
```

```
## Intercept = 10.869
## Slope =  3.285
## Intercept = 10.866
## Slope =  3.287
## Intercept = 10.868
## Slope =  3.285
## Intercept = 10.868
## Slope =  3.285
## Intercept = 10.867
## Slope =  3.285
```

Das oben gezeigte Programm hat 10^5 Runden des Gibbs Samplers gemacht und als Resultat erhalten wir die Schätzung für den Achabschnitt als

$$\hat{\beta}_0 = 10.87$$

Die Bayessche Schätzung für den Koeffizienten der erklärenden Variablen lautet

$$\hat{\beta}_1 = 3.29$$

Ihre Aufgabe

- Vergleichen Sie die Bayessche Schätzung mit der Schätzung aufgrund von Least Squares.
- Da die Daten ursprünglich simuliert waren, kennen wir die wahren Werte diese sind in der nachfolgenden Tabelle gezeigt. Vervollständigen Sie die folgende Tabelle für einen übersichtlichen Vergleich.

Parameter	Wahr	Bayes	LeastSquares
Achsenabschnitt	10.9		
Koeffizient	3.4		